Dr. Babasaheb Ambedkar Open University

(Established by Government of Gujarat)

BAOU Education for All

NAAC ACCREDITED A++ GRADE 3.55/4

BBA/DBA

SEMESTER - 2

BBAMDC203

DBAMDC203

Business Statistics

Message for the Students

Dr. Babasaheb Ambedkar Open (University is the only state Open University, established by the Government of Gujarat by the Act No. 14 of 1994 passed by the Gujarat State Legislature; in the memory of the creator of Indian Constitution and Bharat Ratna Dr. Babasaheb Ambedkar. We Stand at the seventh position in terms of establishment of the Open Universities in the country. The University provides as many as 54 courses including various Certificate, Diploma, UG, PG as well as Doctoral to strengthen Higher Education across the state.

On the occasion of the birth anniversary of Babasaheb Ambedkar, the Gujarat government secured a quiet place with the latest convenience for University, and created a building with all the modern amenities named 'Jyotirmay' Parisar. The Board of Management of the University has greatly contributed to the making of the University and will continue to this by all the means.

Education is the perceived capital investment. Education can contribute more to improving the quality of the people. Here I remember the educational philosophy laid down by Shri Swami Vivekananda:

> ***"We want the education by which the character is formed, strength of mind is Increased, the intellect is expand and by which one can stand on one's own feet".***

In order to provide students with qualitative, skill and life oriented education at their threshold. Dr. Babaasaheb Ambedkar Open University is dedicated to this very manifestation of education. The university is incessantly working to provide higher education to the wider mass across the state of Gujarat and prepare them to face day to day challenges and lead their lives with all the capacity for the upliftment of the society in general and the nation in particular.

The university following the core motto 'स्वाध◌्य◌ाय: परमम ◌् तप:' does believe in offering enriched curriculum to the student. The university has come up with lucid material for the better understanding of the students in their concerned subject. With this, the university has widened scope for those students who are not able to continue with their education in regular/conventional mode. In every subject a dedicated term for Self Learning Material comprising of Programme advisory committee members, content writers and content and language reviewers has been formed to cater the needs of the students.

Matching with the pace of the digital world, the university has its own digital platform Omkar-e to provide education through ICT. Very soon, the University going to offer new online Certificate and Diploma programme on various subjects like Yoga, Naturopathy, and Indian Classical Dance etc. would be available as elective also.

With all these efforts, Dr. Babasaheb Ambedkar Open University is in the process of being core centre of Knowledge and Education and we invite you to join hands to this pious *Yajna* and bring the dreams of Dr. Babasaheb Ambedkar of Harmonious Society come true.

Prof. Ami Upadhyay
Vice Chancellor,
Dr. Babasaheb Ambedkar Open University,
Ahmedabad.

Dr. Babasaheb Ambedkar Open University

(Established by Government of Gujarat)

BBA/DBA

SEMESTER - 2

BBAMDC203

DBAMDC203

BUSINESS STATISTICS

BLOCK-1

BLOCK-2

BLOCK-3

BBA
SEMESTER-2
BUSINESS STATISTICS
BLOCK: 1

Authors’ Name: Dr. Dipak Sanki, Assistant Professor, KBS College, Vapi

Review (Subject): Dr.Vaibhav Shah, Professor, New L J Commerce College, Ahmedabad

Review (Language): Dr. Bhavna Trivedi , Assistant Professor, Dr. BAOU, Ahmedabad

Editor’s Name: Prof. (Dr.) Manoj Shah,
Professor and Director,
School of Commerce and Management,
Dr. Babasaheb Ambedkar Open University,
Ahmedabad.

Publisher’s Name: Dr. Ajaysinh Jadeja,
Registrar,
Dr. Babasaheb Ambedkar Open University
'JyotirmayParisar',Opp.ShriBalaji
Temple,Chharodi,Ahmedabad,382481,
Gujarat,India.

Edition: 2024 (First Edition)

ISBN: 978-93-5598-562-0

978-93-5598-562-0

UNIT-1 INTRODUCTION TO STATISTICS

1.1 Introduction

Welcome to statistics, an exciting field! Consider how frequently you deal with data in your day-to-day activities, such as reading survey findings, watching sports statistics, or monitoring weather updates. Because statistics gives us the means to gather, arrange, and analyse data, it helps us make sense of all this information and enables us to base our decisions on factual information rather than conjecture.

Almost every field is impacted by statistics. It helps businesses to forecast financial trends, comprehend consumer behaviour, and enhance products based on data. Statistics contribute in the study of illnesses, the creation of efficient cures, and the prediction of patient outcomes in the medical field. It reveals trends in education that improve instructional strategies and student experiences. In essence, statistics gives us the information we need to spot patterns, forecast outcomes, and guide decisions wherever data is available.

You will discover different kinds of data, how they are gathered, and how to analyse them as you study this subject. Gaining proficiency in these areas will enable you to evaluate data critically, make sense of it, and confidently face obstacles in the real world.

Statistics is more than just numbers; it is about interpreting the narrative they tell. Although data is frequently complex or unstructured, statistics offers the methods for organizing, interpreting, and deriving meaning from it. For instance, rather than depending solely on a small number of opinions, evaluating the feedback of a big group enables us to identify trends and create evidence-based forecasts about whether a new product would succeed.

The ability of statistics to guide us through uncertainty is another important advantage. From sales projections to weather forecasts, life is full of unknowns. We are able to make educated forecasts about what will happen in the future by using statistical models. To forecast weather patterns, for example, meteorologists use statistical models built on past data. These predictions are extremely helpful, even though they are not flawless, and statistics makes them possible.

Statistics are a very useful tool in the corporate sector. Businesses utilize it for quality control, risk analysis, inventory management, and market research. Statistics are at play if you have ever seen tailored product recommendations when you are shopping online! Businesses make strategic choices that improve customer experiences and advance their bottom line by examining data on consumer preferences.

Developing your ability to think critically is another benefit of studying statistics. Investigating statistical techniques will teach you to challenge data, recognize biases, and notice patterns that others might miss. In today's data-driven world, when careful analysis is frequently necessary to make well-informed decisions, these abilities are crucial.

In addition to learning how to manage data, the course will teach you the importance of the narratives that numbers may convey. Understanding statistics will provide you with a distinct edge in making well-informed and significant judgments, regardless of your career interests—business, science, social sciences, or almost any other sector.

1.2. Objectives

Statistics is like a toolkit for understanding and making decisions based on data. Imagine trying to understand a crowd — it is tough. But with statistics, we can take a small sample and use it to make sense of the whole crowd. Here is what statistics aims to do:

1. Collect Data: The first step is to gather accurate and relevant data. For example, if you are studying students' favourite subjects, you collect responses from a representative group.

2. Summarize Data: Once data is collected, we summarize it using averages, percentages, or frequencies, making it easier to understand and spot trends.

3. Analyse Relationships: Statistics helps us see how different factors are connected, like whether studying more leads to higher test scores. This helps us understand cause-and-effect relationships.

4. Make Predictions: By looking at past data, we can predict future outcomes, like forecasting next year’s weather based on previous trends.

5. Support Decisions: Statistics helps businesses, governments, and individuals make data-driven decisions instead of relying on guesswork.

6. Test Ideas: Through statistical testing, we can confirm or reject hypotheses, like testing if more sleep leads to better test scores.

7. Understand Variability: Not all data is the same, and statistics helps us understand how much things differ, which helps us interpret results accurately.

8. Improve Processes: In fields like healthcare or manufacturing, statistics helps identify areas for improvement and drive better outcomes over time.

1.3 Meaning and Definition of Statistics

1.3.1. Meaning

The science of gathering, evaluating, interpreting, and presenting data is known as statistics. By providing an easily comprehensible and useful summary, it enables us in making sense of vast amounts of information. We may make well-informed decisions and predictions by using statistics to find patterns, trends, and relationships in data. In essence, it transforms data into insightful knowledge.

1.3.2. Definitions of Statistics:

Karl Pearson (British statistician and founder of modern statistics):
"Statistics is the science of counting and measuring, the science of economics, the science of observation, and the science of logic applied to data."

Ronald A. Fisher (English statistician, biologist, and geneticist):
"Statistics is the method of making decisions in the face of uncertainty. It involves the collection, organization, analysis, interpretation, and presentation of data."

George E.P. Box (British statistician):
"Statistics is the art of making inferences from data, especially when we cannot directly observe the phenomena of interest."

John Tukey (American mathematician and statistician):
"The greatest value of a picture is when it forces us to notice what we never expected to see. Statistics is the science of data analysis that helps reveal such unexpected insights."

Sir Francis Galton (English polymath and statistician):
"Statistics is the mathematical study of the frequency and distribution of phenomena in society and nature."

David S. Moore (American statistician and educator):
"Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty."

1.4 Nature and Scope of Statistics

Statistics is a powerful tool that helps make sense of data and guide decisions in various fields, from business to healthcare, and beyond. Its nature is analytical and objective, while its scope is broad and applicable in many areas.

1.4.1 Nature of Statistics

Statistics is both a **science** and a **tool** for decision-making. It deals with data collection, organization, analysis, and interpretation. The nature of statistics includes:

1. **Descriptive Nature**: It entails arranging and summarizing data in order to present it in a comprehensible manner, including tables, graphs, or averages.
2. **Inferential Nature**: It enables us to draw conclusions or forecasts about a broader population from a smaller sample of data.
3. **Quantitative Focus**: The main focus of statistics is numerical data, which enables us to measure relationships and analyse quantities.
4. **Objective:** Statistical methods rely on data and evidence to produce conclusions that are objective and free from personal biases.

5. **Mathematical Foundation**: Statistics analyses data using mathematical ideas and procedures to provide precise and reliable scientific findings.

1.4.2. Scope of Statistics

The scope of statistics is vast, as it can be applied to almost every field of study and industry. Here is a look at some key areas where statistics plays a vital role:

1. **Business and Economics**: Used for market research, forecasting, and making business decisions by analysing consumer trends, sales, and financial data.

2. **Healthcare**: In medicine, statistics help in research, clinical trials, disease control, and health policy planning by analysing patient data and treatment outcomes.

3. **Social Sciences**: In psychology, sociology, and education, statistics helps understand human behaviour, study social trends, and evaluate educational systems.

4. **Agriculture**: It is used to analyse crop yields, livestock performance, and other agricultural data to improve production and efficiency.

5. **Government and Public Policy**: Statistics guides government decisions related to budgeting, planning, population studies, and public health.

6. **Sports**: In sports, statistics helps in analysing player performance, team strategies, and predicting outcomes of games.

7. **Environmental Science**: Used to analyse data on climate change, pollution, and wildlife, helping to form policies for environmental conservation.

1.5 Applications of Statistics in Business

Statistics plays a crucial role in the success of businesses by turning data into valuable insights. Think of it as a tool that helps businesses make informed decisions. Let’s explore how statistics is used across various areas in business:

1. **Market Research:** Businesses must know what customers desire when they introduce new goods or services. In order to forecast purchasing patterns and market trends, statistics analyse data from focus groups, consumer surveys, and other sources. Businesses can use statistical techniques to determine the needs of various client categories and adjust their marketing strategies accordingly.

2. **Financial Analysis:** Effective financial management is crucial in business. Statistics make it easier to monitor earnings, costs, and profits over time. Businesses are better equipped to anticipate future revenues, evaluate possible investments, and create budgets by examining trends. For instance, companies evaluate financial risks and forecast cash flow using statistical techniques like regression analysis.

3. **Quality Control:** In manufacturing, quality control is essential. Production processes are tracked using statistical methods to make sure the final product satisfies quality requirements. Statistical process control (SPC) is a popular technique used by companies to monitor changes in the manufacturing process and spot flaws early. This enhances consumer happiness and cuts down on waste.

4. **Sales Forecasting:** Through the analysis of previous sales data, statistics assist businesses in forecasting future sales. By examining past patterns, companies may predict the demand for goods and services. They are able to control production schedules, optimize inventory, and prevent stockouts and overstocking as a result.

5. **Decision-Making:** When it comes to making strategic decisions, statistics are crucial. For instance, companies frequently employ hypothesis testing to ascertain whether altering their product design or marketing approach will improve outcomes. Confidence intervals are also used to evaluate the risk involved with choices like starting new initiatives or breaking into untapped markets.

6. **Employee Performance and Productivity** Employers use statistical methods to monitor staff performance and productivity. Businesses can determine high performers and areas for development by examining staff statistics, such as sales performance or customer satisfaction ratings. Decisions about training initiatives, incentives, and promotions might also be influenced by this data.

7. **Risk Management:** Whether it is shifting market pricing, natural calamities, or economic downturns, businesses are exposed to hazards. Statistics are useful for detecting possible hazards and quantifying their effects. Businesses can use statistical analysis to create plans to reduce these risks and make well-informed choices about investments, insurance, and crisis management.

8. Businesses would have to guess in the absence of statistics. However, they can use data to make well-informed decisions that result in success and growth. Businesses may forecast future events, lower risks, and increase overall efficiency by analysing data trends. Managers can identify opportunities, resolve issues, and streamline operations with the use of statistics.

❖ Exercise

A. Answer the following questions:

1. Write the meaning of Statistics.
2. Write definition of statistics.
3. Write the nature of statistics.
4. What is the scope of statistics?
5. Write short note on: Applications of Statistics in Business.
6. Write the objective of statistics.

B. Short note on:

1. Discuss the nature of statistics and its significance in analysing data.
2. Explain in detail the scope of statistics and how it applies to various fields like business, healthcare, and agriculture.
3. How do businesses utilize statistics for market research and quality control? Provide examples.
4. What is the role of statistics in financial analysis and sales forecasting? Discuss with suitable illustrations.
5. Explain the advantages of using statistical methods in improving employee performance and productivity.

6. Discuss the applications of statistics in risk management and decision-making processes.
7. Why statistics is called a science of uncertainty? Explain with relevant examples.
8. Write a detailed note on the importance of data collection and analysis in making business decisions.

C. Multiple Choice Questions (MCQs)

1. **What is the primary focus of statistics?**
   a) Generating random numbers
   b) Analysing and interpreting data
   c) Conducting physical experiments
   d) Creating new scientific theories

2. **Which of the following is NOT an objective of statistics?**
   a) Collecting data
   b) Predicting future trends
   c) Proving all theories
   d) Supporting decision-making

3. **Who defined statistics as "the art of making inferences from data"?**
   a) Karl Pearson
   b) Ronald A. Fisher
   c) George E.P. Box
   d) Sir Francis Galton

4. **What is the inferential nature of statistics?**
   a) Organizing data into tables
   b) Drawing conclusions about a population from a sample
   c) Presenting data in graphs
   d) Calculating percentages

5. **In which field are statistical models widely used to forecast weather patterns?**
   a) Medicine
   b) Education
   c) Meteorology
   d) Agriculture

6. **Which of the following is a scope of statistics in healthcare?**
   a) Financial analysis
   b) Quality control
   c) Analysing patient outcomes
   d) Market research

7. **What type of data is primarily analysed in statistics?**
   a) Verbal
   b) Numerical
   c) Visual
   d) Logical

8. **What does the descriptive nature of statistics involve?**
   a) Drawing conclusions about the population
   b) Collecting raw data

c) Summarizing data to make it understandable
d) Predicting future outcomes

9. **What is the primary use of statistics in business decision-making?**
a) Conducting surveys
b) Identifying high performers
c) Reducing guesswork with data-driven decisions
d) Monitoring weather patterns

10. **Which of the following is an application of statistics in agriculture?**
a) Studying human behaviour
b) Analysing crop yields
c) Monitoring employee productivity
d) Forecasting sales

11. **What does statistical process control (SPC) help identify?**
a) Market trends
b) Flaws in manufacturing processes
c) Customer preferences
d) Financial risks

12. **Which of the following fields relies on statistics to study social trends?**
a) Economics
b) Sociology
c) Meteorology
d) Medicine

13. **What is the meaning of statistics?**
a) A method to create data
b) Science of gathering, analysing, and interpreting data
c) A way to guess outcomes
d) Mathematical calculations

14. **How does statistics improve processes in healthcare?**
a) By creating new medicines
b) By identifying improvement areas
c) By predicting weather conditions
d) By studying customer satisfaction

15. **Which of these statements is true about statistics?**
a) It only applies to numerical data.
b) It has limited applications in business.
c) It helps make sense of large amounts of data.
d) It cannot be used to predict future trends.

Answers:

1. **b) Analysing and interpreting data**
2. **c) Proving all theories**
3. **c) George E.P. Box**
4. **b) Drawing conclusions about a population from a sample**
5. **c) Meteorology**
6. **c) Analysing patient outcomes**
7. **b) Numerical**
8. **c) Summarizing data to make it understandable**
9. **c) Reducing guesswork with data-driven decisions**

10. b) Analysing crop yields
11. b) Flaws in manufacturing processes
12. b) Sociology
13. b) Science of gathering, analysing, and interpreting data
14. b) By identifying improvement areas
15. c) It helps make sense of large amounts of data

UNIT-2

COLLECTION OF DATA AND TABULATION

2.1. Introduction:

Data collection is the initial stage in statistics that leads to useful analysis. Data are facts or unprocessed information that we collect to address particular issues or provide answers to particular inquiries. For instance, individual height measurements must be taken if we wish to determine the average height of the pupils in a school. In order to prevent errors in analysis, it is crucial to gather data precisely and consistently. This method guarantees that we have the correct information to work with.

After the data is gathered, it needs to be arranged such that it is simple to understand. Here's where tabulation is useful. Tabulation transforms unstructured material into a structured format by methodically organizing it into rows and columns. It makes complicated data easier to understand and facilitates the quick identification of links, trends, or patterns. It can be difficult to interpret or make inferences from raw data without tabulation. The foundation of each statistical investigation is made up of data gathering and tabulation.

2.2. Collection of Data

Data is basically information collected for a specific purpose. It can be anything we gather to help us understand or analyse something. Think of data as pieces of a puzzle that, when put together, give us a complete picture.

2.2.1. Qualitative Data: The Descriptive Stuff

Qualitative data is all about qualities, characteristics, or descriptions. It's the kind of data that can’t be counted, but can be observed or categorized. For example:

- **Colour of a car** (red, blue, black)
- **Types of food** (fruits, vegetables, meat)
- **Feelings** (happy, sad, excited)

This type of data is often non-numeric. We use words to describe it and group things based on their qualities.

2.2.2. Quantitative Data: The Countable Stuff

Quantitative data, on the other hand, deals with numbers. It's the kind of data that can be measured, counted, and analysed mathematically. For example:

- **Height of a person** (5'8", 6'2")
- **Number of students in a class** (25 students)

2.3. Concept of Data

Data is a collection of facts, figures, and statistics that are used for analysis and decision-making. In statistics, data is classified into two main categories: Variables and Attributes. Understanding these concepts is crucial for organizing and analysing data effectively.

2.3.1. Variable

Definition:

A variable is a quantity that can change or vary from one unit to another. It is typically expressed in numerical form and can take different values depending on the conditions being studied.

Characteristics:

1. Variables are measurable and can be expressed numerically.
2. They allow for statistical investigation of individuals or units.
3. The value of a variable can change, making it useful for studying different characteristics in a population or sample**.**

Examples:

- ✓ Price: The cost of an item can vary from one product to another.
- ✓ Production supply: The number of units produced may change depending on the production process.
- ✓ Height of students: The height of different students can vary, making it a variable.
- ✓ Number of children per family: The number of children in a family is a variable that changes from family to family.

2.3.2. Attribute

Definition:

An attribute refers to a characteristic or property of an individual or unit that cannot be expressed numerically. Attributes help classify or categorize individuals based on their qualities, but they are not measured in numerical terms.

Characteristics:

- ✓ Attributes are non-numerical and are used for classification.
- ✓ They represent qualities or characteristics that describe the units being studied but cannot be quantified.

Examples:

- ✓ Religion: Categories like Hindu, Muslim, Christian, etc., are attributes that classify people based on their religious beliefs.
- ✓ Beauty: Beauty is a subjective quality that can be used as an attribute to describe individuals.
- ✓ Honesty: An individual's honesty is an attribute that cannot be numerically quantified.
- ✓ Marital status: Whether a person is married, single, or divorced, is an attribute representing their relationship status.

2.4. Primary Data and Secondary Data

When we talk about data, we generally classify it into two broad types: **Primary Data** and **Secondary Data**. Let's dive into what these two types are, how they differ, and when you might use them.

2.4.1. Primary Data: First-Hand Information

Primary Data is data that is **collected directly** by the researcher for a specific research purpose or project. It is the **raw data** gathered from original sources, and it has not been processed or interpreted by anyone else yet. Think of it as the **first-hand** information you gather directly from the source.

How is Primary Data Collected?

Primary data can be collected through:

1. **Surveys or Questionnaires** – Asking people specific questions related to the research.
2. **Interviews** – One-on-one conversations with individuals to collect insights.
3. **Experiments** – Conducting experiments to observe and record outcomes.
4. **Observations** – Watching behaviour patterns or events as they happen.
5. **Focus Groups** – Group discussions that explore opinions on a specific topic.

Example of Primary Data:

- **A company conducts a survey** to find out customer satisfaction levels with their new product.
- **A scientist runs an experiment** to measure the effects of a new drug on patients.
- **A researcher observes** how people interact with different types of social media.

Advantages of Primary Data:

- **Specific to Your Study**: It directly addresses the research questions you are interested in.
- **Up-to-Date**: Since you're collecting the data yourself, it is fresh and current.

- **Accurate**: It has not been altered or interpreted by someone else, reducing the chances of errors.

Disadvantages of Primary Data:

- **Time-Consuming**: Collecting primary data can take a lot of time and effort.
- **Expensive**: It may require resources such as money for tools, staff, and logistics.

2.4.2. Secondary Data: Data Collected by Someone Else

Secondary Data refers to data that has already been **collected, processed, and published** by someone else for a different purpose. Instead of going out to collect data yourself, you use existing data that others have gathered, such as published reports, government records, or research articles.

Sources of Secondary Data:

- **Books, Journals, and Articles** – Academic papers, books, or reports published by researchers.
- **Government Reports** – Statistical data provided by government agencies (e.g., census data).
- **Websites** – Data shared by various organizations online.
- **Company Records** – Internal data like sales reports, employee records, etc.

Example of Secondary Data:

- **Using government census data** to analyse population trends in a specific region.
- **Reviewing research articles** that have already studied a particular phenomenon.
- **Analysing sales data** from a company's previous reports.

Advantages of Secondary Data:

- **Less Time-Consuming**: The data is already collected, so you don't have to spend time gathering it yourself.
- **Cost-Effective**: It's usually free or less expensive compared to collecting primary data.
- **Access to a Larger Data Set**: Secondary data may provide access to broader datasets or long-term trends that are difficult to collect on your own.

Disadvantages of Secondary Data:

- **Not Always Relevant**: The data may not perfectly match the specific questions you want to answer.
- **Potential for Errors**: Since the data was collected by someone else, there's a risk it might not be accurate or reliable.
- **Outdated Information**: The data could be old and may not reflect current trends or conditions.

2.5. Different type of scale:

These scales are used to classify and measure data:

- **Nominal Scale**
  - Used for labelling or categorizing without a quantitative value.
  - Examples: Gender (Male/Female), Types of Fruit (Apple, Orange).
- **Ordinal Scale**
  - Represents ordered categories where the exact difference between ranks is not known.
  - Examples: Satisfaction ratings (Satisfied, Neutral, Dissatisfied), Military ranks.
- **Interval Scale**
  - Measures where the difference between values is meaningful, but there is no true zero point.
  - Examples: Temperature (Celsius/Fahrenheit), IQ Scores.
- **Ratio Scale**
  - Similar to an interval scale but includes a true zero, allowing for the calculation of ratios.
  - Examples: Weight, Height, Income.

2.6. Tabulation

Have you ever written down a list to organize information, like a class attendance sheet or survey results? If yes, then you have already taken the first step toward understanding **tabulation**!

Tabulation is a method of presenting raw data in an organized and structured way using rows and columns. It helps transform messy, unorganized data into clear, meaningful tables that are easy to read and interpret.

For example, if you conduct a survey to find out how many students prefer different sports, listing all answers might be confusing. But by organizing the data into a table, you create clarity:

Sport	Number of Students
Football	10
Cricket	15
Basketball	8

This simple table makes the data more understandable at a glance.

Why Is Tabulation Important?

1. **Clarity:** Makes large amounts of data easy to comprehend.
2. **Comparison:** Helps compare different categories effectively.

3. **Analysis:** Serves as a foundation for further statistical analysis like averages, percentages, or trends.

4. **Presentation:** Makes data visually appealing for reports and presentations.

Tabulation is like organizing your wardrobe—when things are neatly arranged, finding what you need becomes much easier!

Objective of tabulations

1. **Organize Data:** Tabulation arranges raw data into rows and columns for clarity.
2. **Simplify Information:** It simplifies complex data for easier interpretation.
3. **Enable Comparison:** Tables make it easy to compare different data categories.
4. **Highlight Trends:** Tabulation reveals patterns and trends in the data.
5. **Save Time:** It reduces the time required for analysing information.
6. **Support Analysis:** Tabulated data is a base for statistical calculations.
7. **Improve Presentation:** It enhances the visual appeal of data in reports or presentations.
8. **Assist Decision-Making:** Tabulation provides summarized insights for decisions.
9. **Minimize Errors:** It reduces errors by organizing data systematically.

2.7. Types of Tabulation with Examples

1. **Simple Tabulation**
    - **Definition:** Organizes data based on a single characteristic or variable.
    - **Example:** Number of students in different classes.

Class	Number of Students
Class 6	30
Class 7	35
Class 8	40

2. **Complex Tabulation**
    - **Definition:** Organizes data based on two or more characteristics or variables.
    - **Example:** Number of students in different classes categorized by gender.

Class	Boys	Girls
Class 6	18	12
Class 7	20	15
Class 8	22	18

Rules of Tabulation

When creating tables to organize data, certain rules must be followed to ensure clarity, consistency, and ease of understanding. Here are the key rules of tabulation:

1. **Clear and Simple Title:**
   - Every table should have a clear and concise title that explains what the table represents.
   - **Example:** "Number of Students in Different Classes."
2. **Proper Heading:**
   - Each column and row should have clear headings to identify the data presented.
   - **Example:** A column header like "Class" or "Gender" should be used.
3. **Consistent Arrangement of Data:**
   - Data should be arranged systematically, either in ascending or descending order, depending on the nature of the data.
   - **Example:** If you are showing age groups, arrange them in increasing order (e.g., 10-15, 16-20, etc.).
4. **Keep Data in a Logical Order:**
   - Organize data in a way that makes sense for comparison or analysis, such as grouping related information together.
   - **Example:** Group data by category or type (e.g., products, regions, or time periods).
5. **Uniform Units of Measurement:**
   - Make sure the units of measurement for the data are consistent across the table.
   - **Example:** If you are displaying sales data, ensure all amounts are in the same currency (e.g., "Sales in USD").
6. **Avoid Clutter:**
   - Keep the table clean and avoid unnecessary details that can make the table difficult to read.
   - **Example:** Use minimal text in cells and avoid adding extra rows or columns that are not needed for analysis.
7. **Use of Footnotes (if necessary):**
   - If certain data requires further explanation or clarification, footnotes should be added at the bottom of the table.
   - **Example:** "Note: Data for Q4 2023 is estimated."

8. **Proper Alignment of Data:**
   - Numbers should be aligned to the right, text should be aligned to the left, and headings should be centred for clarity.
   - **Example:** Right-align numbers for easy comparison.
9. **Avoid Overlapping Data:**
   - Ensure that data in rows and columns does not overlap, which can cause confusion.
   - **Example:** Don't cram too much information into a single cell.
10. **Use of Appropriate Number of Columns and Rows:**
    - Avoid making the table too large or small. Make sure the number of columns and rows is just enough to present the data without overcrowding.
    - **Example:** If you have a lot of data, consider breaking it into smaller, more manageable tables.

2.8. Frequency Distribution and Its Types

Frequency distribution is a way to organize and summarize data, showing how often each value or range of values occurs in a data set. It is essential in statistics because it allows for easy analysis and understanding of large data sets. Let's explore the different types of frequency distributions:

2.8.1. Discrete Frequency Distribution

Definition:
In a discrete frequency distribution, the data consists of distinct, separate values, often whole numbers (integers), where each value occurs a certain number of times.

Example no. 1: Prepare a frequency distribution from the data:

18,18,17,15,14,14,15,18,18,15,14,15,16,17

Solution:

X	Tally marks	frequency
14	III	03
15	IIII	04
16	I	01
17	II	02
18	IIII	04
Total		14

2.8.2. Continuous Frequency Distribution

Definition:
A continuous frequency distribution deals with data that can take any value within a given range, typically representing measurements that can be divided into smaller increments. These values are usually grouped into intervals.

Example no. 2

Let's say you measure the heights (in cm) of 15 students and obtain the following data:
150, 155, 160, 165, 155, 170, 175, 160, 165, 160, 150, 155, 170, 160, 155

Solution: We can group these data into intervals and count the frequency of each interval.

Height Range (cm)	Frequency (f)
150-159	5
160-169	6
170-179	4

Interpretation:

- 5 students have heights between 150 and 159 cm.
- 6 students have heights between 160 and 169 cm, and so on.

2.8.3. Cumulative Frequency Distribution

Definition:
A **cumulative frequency distribution** shows the running total of frequencies. It helps in understanding how many values fall below a certain point in a dataset.

Example no 3: Calculate the cumulative frequency from the previous example and construct a **cumulative frequency distribution**.

Solution: Let's take the same **height data** from the previous example and construct a **cumulative frequency distribution**.

Height Range (cm)	Frequency (f)	Cumulative Frequency
150-159	5	5
160-169	6	11
170-179	4	15

Interpretation:

- The cumulative frequency tells us that:
    - 5 students have a height of 159 cm or less.
    - 11 students have a height of 169 cm or less.
    - 15 students have a height of 179 cm or less.

2.8.4. Bivariate Frequency Distribution

Definition:

A **bivariate frequency distribution** involves two variables. It shows the relationship between two different variables by displaying the frequency of each combination of values.

Example no 4: Suppose we have data on the age and smoking habits of a group of people, and we want to create a bivariate frequency table to see how age and smoking habits are related.

Person	Age Group	Smokes (Yes/No)
1	18-25	Yes
2	18-25	No
3	26-35	Yes
4	26-35	No
5	36-45	Yes
6	36-45	No
7	46-60	Yes
8	46-60	No
9	18-25	Yes
10	18-25	No

Solution

Bivariate Frequency Table:

Age Group	Smokes (Yes)	Smokes (No)	Total
18-25	**2**	**2**	**4**
26-35	**1**	**1**	**2**
36-45	**1**	**1**	**2**
46-60	**1**	**1**	**2**
Total	**5**	**5**	**10**

❖ **Exercise**

Short Questions:

1. Define data and explain its types.
2. What is the difference between qualitative and quantitative data?
3. Explain primary data with examples.
4. Describe secondary data with suitable examples.
5. What are the different types of data scales used in statistics?
6. What is tabulation, and why is it important in data analysis?
7. Explain the term frequency distribution and its significance.
8. Differentiate between ungrouped and grouped frequency distribution.
9. List the different types of tabulation with examples.
10. What is a frequency distribution table? How does it help in data analysis?

Long Questions:

1. Discuss the process of data collection. Highlight the key differences between primary and secondary data.

2. Explain the concept of qualitative and quantitative data. Provide examples of each and discuss how they can be used in statistical analysis.

3. Describe the various types of scales used in data collection. How do they affect the analysis and interpretation of data?

4. What is the importance of tabulation in organizing data? Explain different types of tabulation with relevant examples.

5. Explain the process of creating a frequency distribution. Discuss the different types of frequency distributions and their uses in data analysis.

6. Compare and contrast primary and secondary data. Discuss the advantages and disadvantages of using each type of data for research.

7. Elaborate on the concept of frequency distribution. Explain how it can be used to summarize large sets of data. Discuss the various types of frequency distributions.

8. How do data scales influence statistical analysis? Provide an in-depth explanation of nominal, ordinal, interval, and ratio scales.

9. Create a frequency distribution table for a given data set, and explain how it helps in analysing and interpreting the data.

UNIT-3 DIAGRAMS & GRAPHS

2.1. Introduction of Diagrams and Graphs

Consider yourself attempting to interpret data, whether it be numbers, connections, or patterns. Reading through rows and columns could be an option, but it can be stressful and confusing. Graphs and diagrams can help with that! They make complex data much easier to interpret by giving us a clear, visual representation of the information.

Information is represented by shapes, lines, bars, and colours in diagrams and graphs, which draw attention to patterns and trends that could otherwise be overlooked. Every kind of graph, from pie charts and flow diagrams to bar charts and line graphs, has a specific function. For example, pie charts depict proportions within a whole, line graphs indicate changes over time, and bar charts are excellent for comparing numbers.

Diagrams and graphs help us swiftly interpret information and make well-informed decisions by graphically arranging data. They are essential in making complicated subjects understandable so that everyone can easily and clearly examine data, whether in business, education, science, or daily life.

2.2. Objectives of Diagrams and Graphs

The objectives of diagrams and charts are to transform data into a powerful tool for learning, sharing, and decision-making; it is essential for any field to make diagrams and graphs.

1. **Simplify Complex Data:** By displaying vast amounts of data in a visual way, diagrams and graphs help us swiftly understand even the most complex material.

2. **Highlight Patterns and Trends:** They help us spot trends, such as increases or decreases over time, and identify patterns that would be hard to see in raw data alone.

3. **Enhance Comparisons:** By displaying multiple data points side by side, diagrams and graphs make it easier to compare different categories, groups, or time periods.

4. **Support Decision-Making:** With clearer data insights, we can make well-informed choices, whether we're deciding on a business strategy, a research direction, or everyday matters.
5. **Improve Data Retention:** Visuals like diagrams and graphs are often more memorable than plain text or numbers, helping us remember information longer.
6. **Encourage Audience Engagement:** Interactive or well-designed visuals keep people interested and make data presentations more engaging, drawing their attention to important details.
7. **Illustrate Relationships and Correlations:** Graphs show us connections between variables, like the relationship between study time and test scores, which can deepen our understanding of cause and effect.
8. **Simplify Communication of Findings:** Diagrams and graphs help us explain research findings or complex concepts more effectively, making presentations and reports easier to follow.
9. **Increase Accessibility:** For those who struggle with numbers, visual representations of data can make information more accessible and inclusive.
10. **Save Time in Analysis:** Visual tools enable quick overviews, allowing us to analyse large datasets faster and focus on the most critical insights without getting lost in the details.

2.3. Significance of Diagrams and Graphs

Graphs and diagrams are vital tools that transform how we perceive and engage with information, and they are not only for making data seem nice. Consider attempting to evaluate pages of data without any illustrations. It would take a lot of effort and time! Diagrams and graphs help us understand information and uncover insights that may be obscured in tables or plain text by transforming data into visual formats.

The ability of graphs and diagrams to clarify complex data is one of its main benefits. Our brains can easily get overloaded with knowledge when confronted with large volumes of it. However, visuals compress this information into well-structured, easily assimilated chunks. A pie chart immediately displays proportions, assisting us in understanding the makeup of a whole, but a line graph, for instance, can plainly illustrate patterns over time, such as changes in sales or an increase in temperature.

Furthermore, it is much simpler to identify patterns and relationships when using diagrams and graphs. A scatter plot, for example, can show whether two variables move in tandem or in opposition to one another, indicating correlations or cause-and-effect linkages. This helps us find connections that might otherwise go unnoticed, which is crucial for research and decision-making.

In comparisons, graphs and diagrams are also quite important. Examine a bar graph that displays the relative popularity of several goods. Instead of looking through lists of figures, we can quickly determine which objects are performing the best, which makes comparisons much easier. This is particularly beneficial in the business world, where being aware of these variations can result in more effective plans and wiser financial decisions.

Visual aids can greatly improve communication when it comes to sharing information with others. Consider a scientific presentation, business conference, or classroom. Diagrams and graphs make difficult concepts easier for audiences to understand. Without requiring in-depth technical understanding, they may help people from a variety of backgrounds interact with the data by making talks more impactful and understandable.

Lastly, graphs and diagrams improve the retention of material. Visuals are easier for individuals to recall than just words or numbers, according to research. We're more likely to remember the knowledge we get when we use these tools, which can aid us in making wise decisions down the road.

The significance of diagrams and graphs lies in their ability to simplify comparisons, reveal hidden patterns, improve memory retention, facilitate communication, and make data easier to interpret. They are more than simply images; they are essential tools for learning, making decisions, and communicating effectively. They influence how we evaluate and respond to information on a daily basis.

2.4. Importance of Visual Presentation of Data

The visual presentation of data is not just a stylistic choice; it is a crucial part of how we process and understand information. When data is presented in a visual format, it becomes easier to interpret, analyse, and share. The following are reasons why it is so important:

1. **Enhances Understanding**: Numerical or textual raw data can be daunting and challenging to process. Charts, graphs, and diagrams are examples of visual presentations that simplify and make difficult-to-understand datasets easier to interpret. This eliminates the need to sort through mountains of data in order to swiftly locate important information.

2. **Speeds Up Decision-Making**: Finding patterns, trends, and outliers is far simpler when data is displayed visually. We are able to make decisions more quickly and efficiently because of this rapid comprehension. For instance, executives can make quicker strategic decisions based on the obvious upward or decreasing patterns in a line graph that displays a company's revenue growth over time.

3. **Reveals Insights and Patterns**: Data visualization makes connections and patterns that might not be immediately apparent easier to see. For example, a pie chart displays the percentage of each category in the total, whereas a scatter plot might illustrate correlations between variables. Compared to raw data, these visual cues provide us a deeper understanding of linkages and trends.

4. **Improves Retention**: According to studies, information presented visually is far more memorable than information delivered in text or numerical form. Because they are more captivating and memorable, diagrams, graphs, and infographics help us remember information for longer periods of time and refer to it later.

5. **Facilitates Communication**: Visual data presentations facilitate the communication of findings to an audience, whether you are presenting in a report, meeting, or classroom. They make difficult concepts easy to understand, even for those who have little or no prior knowledge of the

subject. For instance, financial estimates presented as graphs rather than long discussions greatly improve the clarity of a business pitch.

6. **Supports Comparisons**: Comparing various data sets is made considerably simpler with the use of visualizations. For example, side-by-side pie charts or bar charts let users compare numbers and categories immediately. When comparing various product characteristics, examining sales statistics, or assessing performance, this is crucial.

7. **Increases Engagement**: People are attracted to images by nature. Dry data are less captivating and engaging than charts, graphs, and infographics. Their ability to swiftly grab and hold attention guarantees that the audience will stay focused on the point you are making.

8. **Clarifies Complex Information**: Certain ideas are intrinsically complicated and challenging to convey with text alone. These concepts are made clearer by visual data displays, which divide them into easier-to-understand parts. For instance, a Venn diagram illustrates the link between various data sets, whereas a flowchart might describe a procedure step-by-step.

9. **Aids in Analysing Trends Over Time**: For monitoring changes over time, visual aids like bar charts and line graphs are ideal. Visual presentations let us see trends, cycles, and growth patterns that might not be immediately apparent in a table of figures, whether we're tracking environmental data, corporate success, or personal development.

10. **Promotes Data-Driven Decisions**: Data becomes easier to obtain with clear graphics, allowing stakeholders to base their decisions on the facts. Data is easier to understand and more rational decisions are made as a result of visual displays rather than intuition or presumptions.

2.5. General Rules for Constructing Diagrams

Creating clear and effective diagrams is a skill that improves with practice. Following a few essential guidelines can make diagrams more appealing, accurate, and easier to understand. Here are some general rules to keep in mind:

1. **Neat and Attractive Presentation**: A diagram should be visually appealing and neatly drawn. This helps capture the viewer's attention and makes the information more engaging.

2. **Accuracy and Proportion**: Ensure that all measurements in the diagram are accurate and proportionate. The correct dimensions of shapes and figures enhance clarity and prevent misinterpretation of data.

3. **Appropriate Size**: The size of the diagram should be proportional to the size of the paper or screen. Diagrams that are too small or too large may lose detail or appear cluttered.

4. **Clear and Concise Heading**: Each diagram should have a brief, suitable title that clearly describes the content. A good heading provides context and helps the reader understand what the diagram represents.

5. **Mention the Scale**: Always include the scale used in the diagram. This allows viewers to interpret the size and measurements accurately, making comparisons easier.

6. **Use of Drawing Instruments**: Diagrams should be drawn with proper tools to ensure neatness and precision. This reduces errors and enhances the professionalism of the presentation.
7. **Include an Index or Legend**: An index or legend helps identify different components in the diagram, such as colours, shapes, or patterns, allowing the reader to interpret the information correctly.
8. **Add a Footnote**: A footnote at the bottom can provide additional context, such as the source of data or special notes. This can be useful for clarifying specific details without cluttering the main diagram.
9. **Economy of Effort and Resources**: Aim for simplicity and efficiency. Avoid excessive detail and unnecessary elements to save time, effort, and resources in the creation of diagrams.

Following these rules helps in producing diagrams that are not only informative but also visually effective, making data presentation clearer and more professional.

2.6. Types of Diagrams

Here is a look at different types of diagrams, each suited for presenting data in a unique way. Diagrams help simplify complex information, making it easier to compare, analyse, and understand data at a glance. Choosing the right type of diagram allows us to highlight specific trends, distributions, or relationships within the data. Let's explore some of the most common types of diagrams and how each one can be used to effectively communicate information.

1. **Line Diagram**:

A line diagram (or line graph) is used to represent data points connected by straight lines, making it ideal for showing trends over time. In this example, we see values plotted for each year from 2020 to 2023, connected by a line. The steady upward trend reveals growth over the years, making it easy to visualize increases or decreases over time.

Example:

Consider the following data for the number of cars sold in a dealership over 6 months:

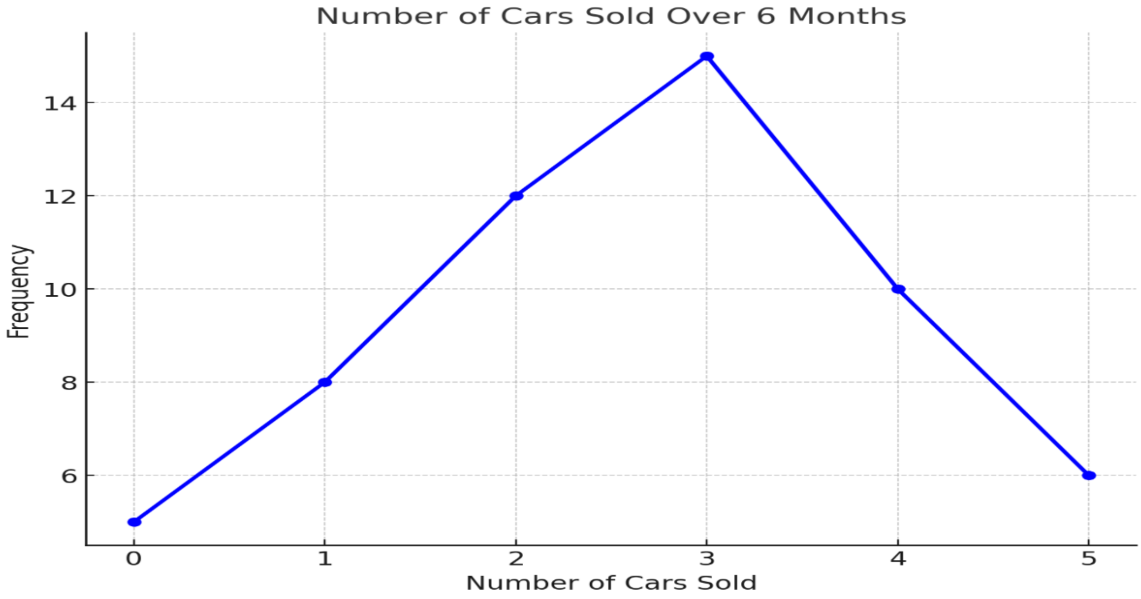
Number of Cars Sold (X)	**Frequency (Y)**
0	5
1	8
2	12
3	15
4	10
5	6

Steps to create the line chart:

1. **Label the Axes:**
   - The **X-axis** represents the **Number of Cars Sold (0, 1, 2, 3, 4, 5)**.
   - The **Y-axis** represents the **Frequency** (5, 8, 12, 15, 10, 6).
2. **Plot the Points:**
   - For each value of the number of cars sold (X), plot the corresponding frequency (Y) as a point on the chart.
3. **Connect the Points:**
   - After plotting the points, connect them with a straight line to show the trend of car sales over the months.

Now, let us create a line chart based on the above data.

Here is a generated chart below:



Here is the line chart representing the number of cars sold over 6 months. The X-axis shows the number of cars sold, and the Y-axis shows the frequency. Each point on the line indicates the frequency of sales for each number of cars sold, and the line connects these points to display the trend.

This chart helps visualize how the frequency of car sales changes with the number of cars sold over the given months.

2. Simple Bar Diagram: A simple bar diagram is straightforward, using single bars to represent data. Each bar stands alone and shows the magnitude of a category for a particular year. For instance, in the example above, each bar represents a single value for each year, allowing a quick comparison of values across time.

Example: Represent the following data by a Simple Bar Diagram.

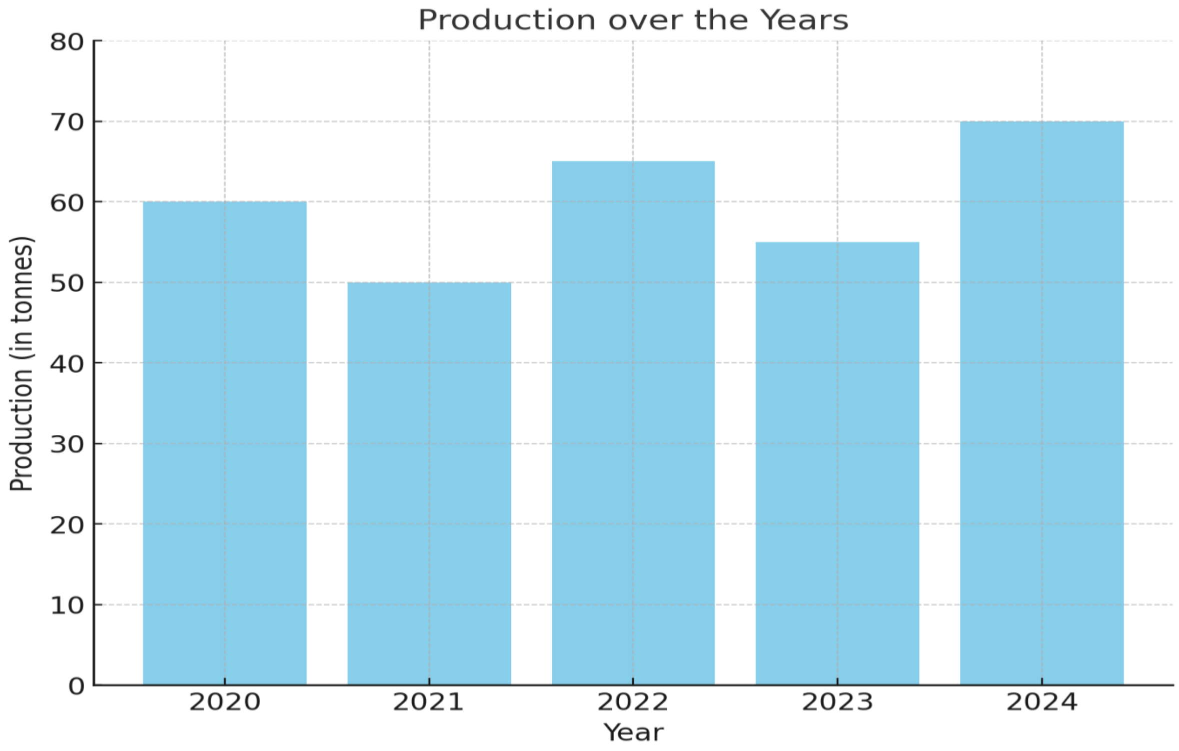
Year	Production (in tonnes)
2020	60
2021	50
2022	65
2023	55
2024	70

Solution:

1. Draw a horizontal axis (X-axis) and a vertical axis (Y-axis). Label the X-axis with the years (2020 to 2024) and the Y-axis with the production values, ranging from 0 to 80 (in increments of 10).
2. Now, plot the bars for each year based on the production values:
    - For 2020, the production is 60 tonnes, so draw a bar up to the 60 marks.
    - For 2021, the production is 50 tonnes, so draw a bar up to the 50 marks.
    - For 2022, the production is 65 tonnes, so draw a bar up to the 65 marks.
    - For 2023, the production is 55 tonnes, so draw a bar up to the 55 marks.
    - For 2024, the production is 70 tonnes, so draw a bar up to the 70 marks.

Now, the bar diagram is complete, and you can easily compare the production across the years.

Here is the diagram based on this data:



Here is the Simple Bar Diagram representing the production data from 2020 to 2024. The bar heights show the production values for each year, making it easy to compare them visually.

3. Multiple Bar Diagram: In a multiple bar diagram, two or more bars are grouped together for each category or year. This setup allows for a side-by-side comparison of different categories within the same timeframe. In the example, each year has two bars representing "Category A" and "Category B," so we can easily see how the categories compare within each year.

Example:

Draw a multiple bar diagram for the following data:

Year	Revenue (in lakhs)	Expenditure (in lakhs)
2020	250	150
2021	300	180
2022	350	220
2023	400	250

Step-by-Step Explanation:

1. **Understand the Data**: The data shows the revenue and expenditure of a company for the years 2020 to 2023. We will represent this data using a **multiple bar diagram**.
2. **Set up the Axes**:

- The **x-axis** (horizontal axis) will represent the years: 2020, 2021, 2022, and 2023.
- The **y-axis** (vertical axis) will represent the amount in lakhs, ranging from 0 to 400 in this case.

3. **Plot the Bars**:

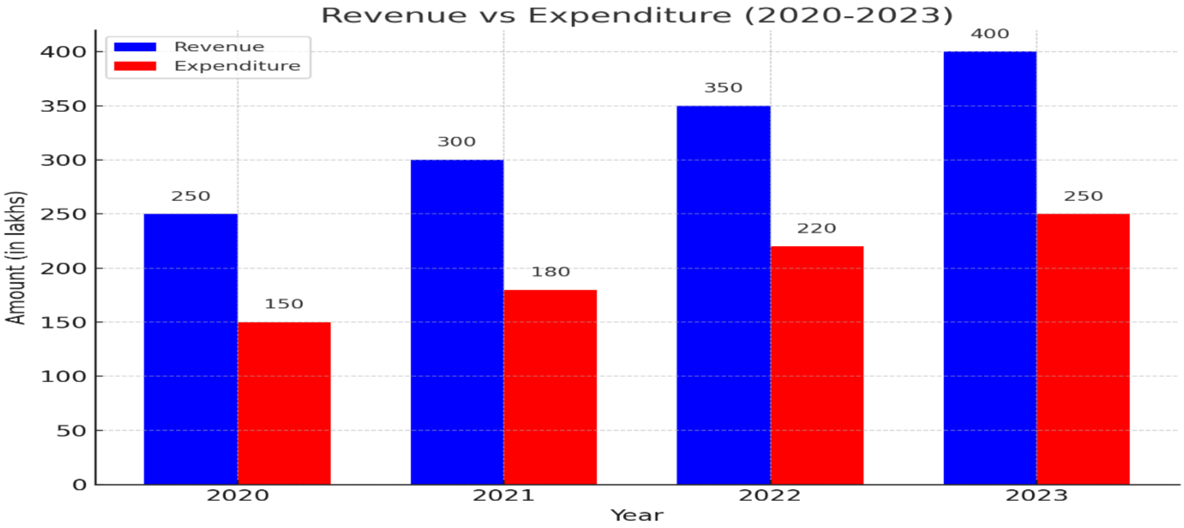
   - For each year, you will have two bars: one for revenue and one for expenditure.
   - The height of the first bar (representing revenue) will correspond to the value in the "Revenue" column.
   - The height of the second bar (representing expenditure) will correspond to the value in the "Expenditure" column.

4. **Colour Code the Bars**:

   - You can use different colours for each bar, for example, **blue for revenue** and **red for expenditure**, to make it easier to differentiate between the two.

5. **Label the Bars**:

   - Label each bar at the top with the exact values (in lakhs) for better clarity.



Here is the multiple bar diagram for the given data, comparing **Revenue** and **Expenditure** for the years 2020 to 2023. The bars are color-coded for easy identification: **blue for revenue** and **red for expenditure**. The exact values are also labelled at the top of each bar.

4. Sub-divided: A sub-divided bar diagram (or stacked bar chart) shows parts of a whole within each bar. Each bar is split into segments that represent different parts, making it easy to see the contribution of each part to the total. For example, each bar has two segments: "Main" and "Subdivision." This allows viewers to see both the total and the contribution of each segment for every year.

Example:

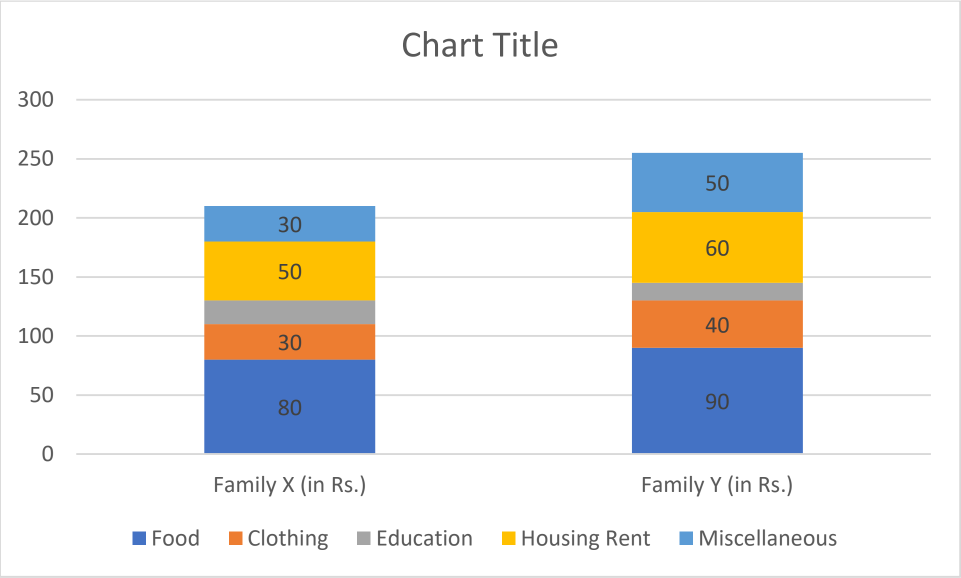
Represent the following data for monthly expenditure of two families using a sub-divided bar diagram.

Expenditure Items	Family X (in Rs.)	Family Y (in Rs.)
Food	80	90
Clothing	30	40
Education	20	15
Housing Rent	50	60
Miscellaneous	30	50

Steps to Draw a Sub-divided Bar Diagram:

1. **Choose the Scale:** Choose a scale, say 1 unit = 10 Rs., to keep the bars proportional to the expenditure.
2. **Draw the Bars:**
   - Each expenditure item will have two bars, one for Family X and one for Family Y.
   - The total height of each bar represents the total expenditure for that family on the respective item.
3. **Sub-divide the Bars:**
   - For Family X: Start from the base, divide the bar into sections representing the expenditure categories: Food, Clothing, Education, Housing Rent, and Miscellaneous.
   - For Family Y: Do the same.

   Label the Sections: Each section should be labelled with the amount for that category, and use different colours or patterns for each section to distinguish them clearly.



5. Pie Diagram: A **pie diagram** (or pie chart) is a circular graph divided into slices to represent numerical proportions. Each slice's size corresponds to a percentage of the total. It is commonly used to show parts of a whole, such as market share, budget allocation, or survey results.

Draw a pie diagram for the following data of production of rice in quintals by various countries.

Country	Production of Rice (in Quintals)
China	120
India	85
Indonesia	60
Brazil	40
Thailand	30

Solution:

1. **Step 1: Calculate the total production of rice.** Add the production values of all countries:

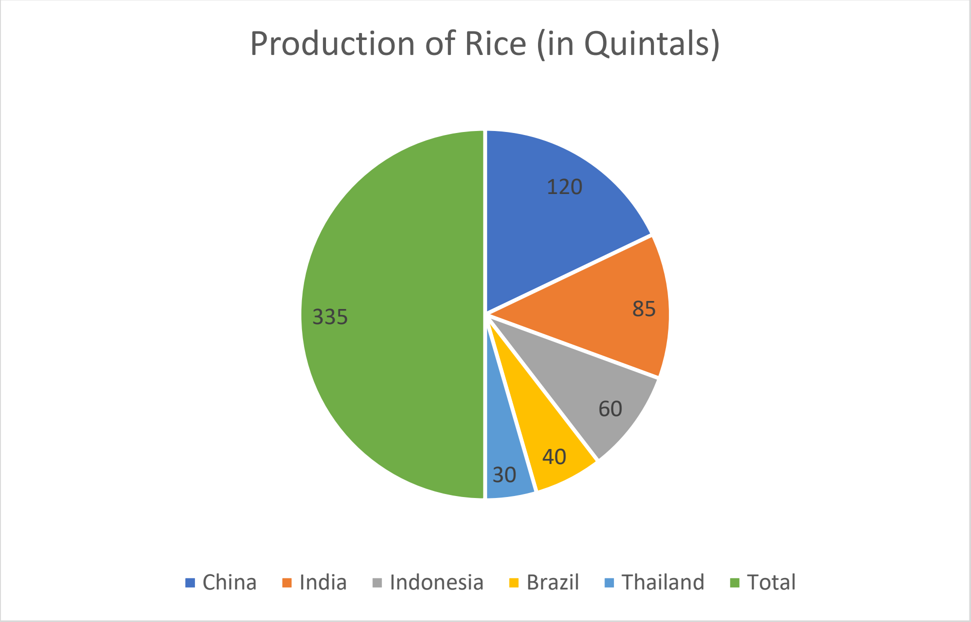
Total production=120+85+60+40+30=335

2. **Step 2: Convert the production values into degrees.** Since the total degrees in a circle is 360°,
3. **Step 3: Summarize the data.**

Country	Production of Rice (in Quintals)	Degrees
China	120	128.57
India	85	91.57
Indonesia	60	64.78
Brazil	40	43.28
Thailand	30	32.14
Total	335	**360**

Step 4: Draw the Pie Chart.

Here is the visual representation of the pie chart based on the calculated degrees for each country. Here is generated the pie chart:



2.7. Graph

Let us dive into graphs! A graph is simply a visual way of presenting data. It makes understanding numbers much easier because it uses shapes and lines to represent information. Graphs are often more appealing and simpler to understand than just reading through tables or figures.

Today, we will cover some of the most popular types of graphs:

1. Histogram

A histogram is a bar graph that shows the distribution of a set of data. Each bar represents the frequency (or count) of data points within a certain range. The bars are usually touching, indicating that the data is continuous. It's great for visualizing how data is spread over intervals, like ages, test scores, or temperature ranges.

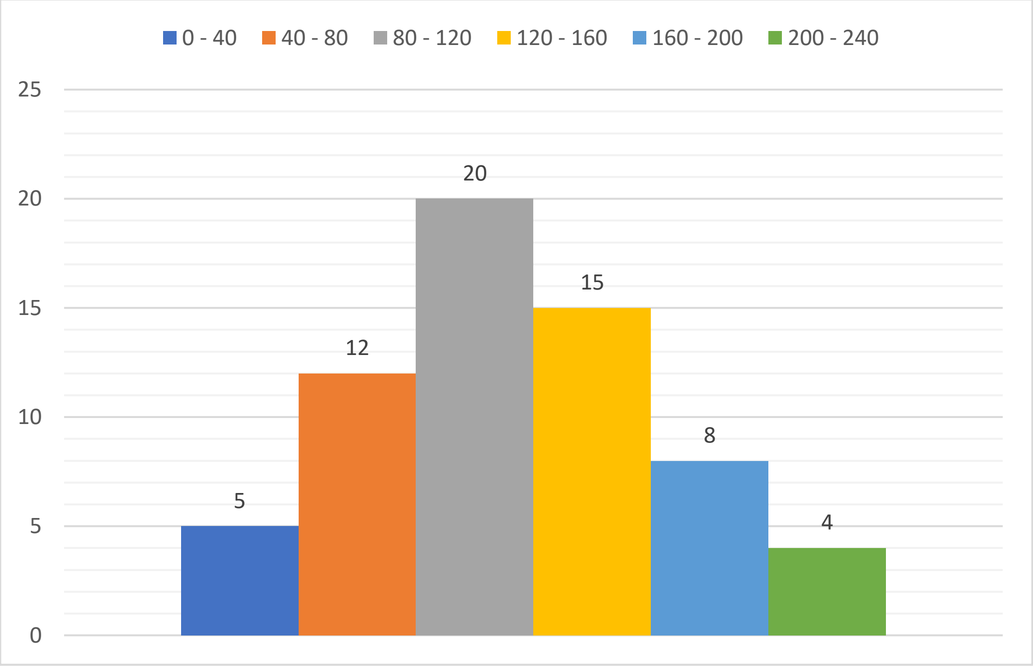
Example:

Draw a histogram for the following data:

Daily Wages	**Number of Workers**
0 - 40	5
40 - 80	12
80 - 120	20
120 - 160	15
160 - 200	8
200 - 240	4

Steps to Draw the Histogram:

1. **Identify the intervals** (the "Daily Wages" column) — these are your x-axis values.
2. **Plot the frequency** (the "Number of Workers") — these are your y-axis values.
3. **Draw bars** — for each interval, the height of the bar corresponds to the number of workers.



2. Frequency Polygon

A frequency polygon is like a line graph that connects the midpoints of the top of each bar in a histogram. It helps us see the shape of the data's distribution, such as whether it is symmetrical, skewed, or has multiple peaks. This graph is useful for comparing multiple datasets.

Example:

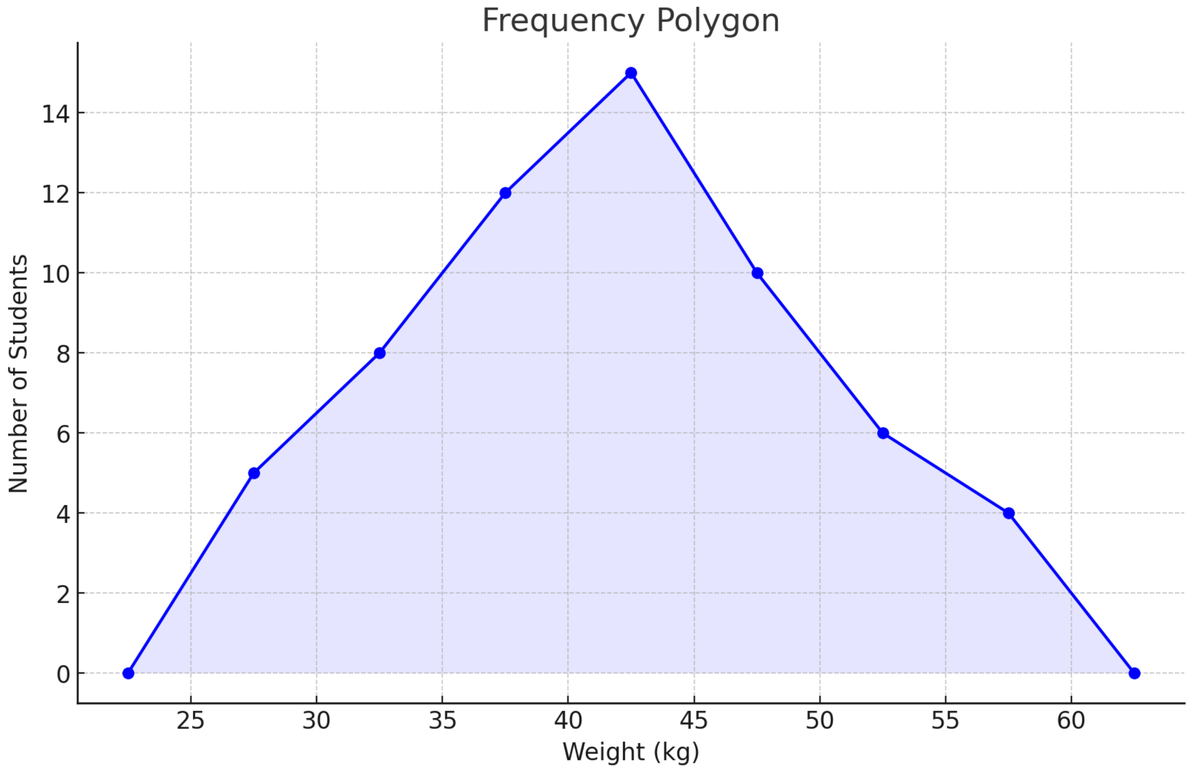
We have the following data on the number of students in various weight ranges:

Weight (in kg)	Number of Students
25 - 30	5
30 - 35	8
35 - 40	12
40 - 45	15
45 - 50	10
50 - 55	6
55 - 60	4

Steps to Draw a Frequency Polygon:

1. **Calculate the midpoints** for each weight class. The midpoint is the average of the lower and upper boundaries of each class.
    - Midpoint of 25-30 = (25 + 30) / 2 = 27.5
    - Midpoint of 30-35 = (30 + 35) / 2 = 32.5
    - Midpoint of 35-40 = (35 + 40) / 2 = 37.5
    - Midpoint of 40-45 = (40 + 45) / 2 = 42.5
    - Midpoint of 45-50 = (45 + 50) / 2 = 47.5
    - Midpoint of 50-55 = (50 + 55) / 2 = 52.5
    - Midpoint of 55-60 = (55 + 60) / 2 = 57.5
2. **Plot the points**: Each midpoint represents the x-coordinate, and the corresponding number of students is the y-coordinate.
    - (27.5, 5)
    - (32.5, 8)
    - (37.5, 12)
    - (42.5, 15)
    - (47.5, 10)
    - (52.5, 6)
    - (57.5, 4)
3. **Connect the points**: Draw straight lines between each plotted point.
4. **Extend the graph**: You can add points for the classes before 25 and after 60 (optional) with a frequency of 0 to close the graph. For example:
    - Point (22.5, 0) for the class before 25
    - Point (62.5, 0) for the class after 60

This will give you a smooth line connecting all the points, which is the frequency polygon. Now, let us visualize it! Here is the graph generated as an example :



Here is the frequency polygon based on the data you provided! The graph shows the distribution of the number of students across different weight ranges. The line connects the midpoints of each weight class, giving us a clear visual representation of how the frequencies change

❖ **Exercise:**

1. What are diagrams and graphs?
2. State the objectives of using diagrams and graphs.
3. Explain the significance of diagrams and graphs in data analysis.
4. Discuss the importance of visual presentation of data.
5. What are the general rules for constructing effective diagrams?
6. Describe the key features of a line diagram.
7. What are the different types of diagrams used for data presentation?
8. Discuss the benefits of using bar charts for data comparison.
9. Explain how pie charts are useful in representing proportions.
10. How do diagrams and graphs make data more accessible to a wider audience?
11. Describe the difference between a line diagram and a bar diagram.

UNIT-4 MEASURES OF CENTRAL TENDENCY

4.1 Introduction

In our daily lives, we often deal with numbers—marks in exams, expenses in a month, or even the temperatures recorded throughout a week. With so much data around us, it can feel overwhelming to understand the overall picture. How can we summarize all these numbers to get an idea of the "average" or "typical" value? This is where the concept of **measures of central tendency** becomes essential.

Measures of central tendency provide us with a single value that represents the entire dataset, giving us a sense of its central or most common characteristic. These measures help us answer questions like: *What is the average score of the class? How much does a typical person spend on groceries?* or *What is the usual height of people in a particular region?*

This chapter introduces the three main measures of central tendency: **mean**, **median**, and **mode**. Each of these has a unique way of summarizing data and is used depending on the type of data we're analysing and the purpose of our study. For example:

- The **mean** gives us an arithmetic average, which works best for evenly distributed data.
- The **median** identifies the middle value, making it useful when data has extreme outliers.
- The **mode** highlights the most frequently occurring value, ideal for categorical or repetitive data.

Understanding these concepts not only strengthens our analytical skills but also equips us to make better decisions based on data. Whether we're conducting research, solving real-world problems, or simply trying to understand trends, measures of central tendency serve as the foundation for exploring and interpreting data in meaningful ways.

Through this chapter, you'll explore the importance of these measures, how to calculate them, and when to apply each one effectively. By the end, you'll see how these simple tools can transform complex datasets into easy-to-understand insights.

4.2 Measures of Central Tendency

4.2.1. Meaning

Measures of central tendency are statistical tools that summarize a large dataset into a single value, representing the "centre" or typical characteristic of the data. Essentially, these measures help identify a point around which the data is clustered. They provide a quick and easy way to understand the overall trend of the data without having to examine each individual value.

For instance, in a class test, knowing everyone's marks can be time-consuming, but calculating the average score gives a clear idea of the class's overall performance. Measures like the **mean**, **median**, and **mode** are the most common tools to determine central tendency, each with its unique way of describing the data's central value.

4.2.2. Objectives

1. **Simplifying Data**: The primary goal is to simplify complex datasets. Instead of analysing hundreds of values, a single representative number gives an overview of the data.
2. **Comparing Datasets:** Measures of central tendency allow us to compare two or more datasets effectively. For instance, we can compare the average scores of two classes to see which performed better.
3. **Describing Data Characteristics**: These measures provide insights into the overall behaviour of the data. For example, they can show whether most values are close to the centre or spread far apart.
4. **Facilitating Decision-Making**: Decision-making in fields like business, healthcare, and education often relies on understanding central values. For example, businesses analyse average sales to predict future trends and make informed choices.
5. **Identifying Patterns**: Measures like the mode help identify trends, such as the most popular product in a market or the most common age group of customers.
6. **Handling Real-Life Problems**: In everyday situations, these measures help us summarize and interpret data effectively. From calculating average expenses to understanding survey results, measures of central tendency are practical tools for problem-solving.

By learning about the meaning and objectives of these measures, you’ll see how they are not just mathematical concepts but essential tools for analysing and understanding the world around us.

4.3 Arithmetic Mean

The mean, often referred to as the average, is one of the most commonly used measures of central tendency. It is calculated by adding up all the values in a dataset and then dividing by the total number of values.

Here's how to calculate the mean:

1. **Add Up the Data:** Start by summing all the values in the dataset.

2. **Divide by the Total Number of Values**: Next, divide the sum by the number of values in the dataset.

For example, let’s calculate the mean for this set of numbers:
5,8,12,15,20

1. Sum the Data:
   5+8+12+15+20=605
2. Divide by the Number of Values:
   There are 5 values, so we divide 60 by 5:
   60/5=12

So, the mean of this dataset is 12.

Why use the Mean?

- The mean is useful when we have data that is evenly distributed and there are no extreme values (outliers) that can skew the result. It gives a good overall summary of the data.
- It’s easy to compute and understand, making it helpful for many practical situations, such as calculating the average score of students in a class or the average sales of a store in a month.

4.3.1. Merits and Demerits of the Mean

Merits of the Mean:

1. **Simple and Easy to Calculate**: The mean is straightforward to compute, as it involves simple addition and division. It’s often the first choice when dealing with a dataset.
2. **Uses All Data Points**: The mean takes every value in the dataset into account, so it provides a comprehensive summary of the data.
3. **Widely Understood**: Since the mean is commonly used and well-understood, it’s easy to communicate and compare data across different groups or studies.
4. **Mathematically Useful**: The mean is often used in further statistical calculations and is the basis for many statistical techniques like standard deviation and variance.
5. **Best for Symmetric Distributions**: When the data is evenly distributed (without outliers), the mean provides an accurate representation of the central value.

Demerits of the Mean:

1. **Sensitive to Outliers**: The biggest drawback of the mean is that it can be heavily influenced by extreme values (outliers). For example, if most students in a class scored between 50-60, but one student scored 100, the mean could be distorted and might not represent the "typical" score.
2. **Does Not Reflect Skewed Data Well**: In datasets with a skewed distribution, the mean may not give a true reflection of the central tendency. For instance, in income data where a few individuals earn extremely high salaries, the mean income could be much higher than what most people earn.

3. **Not Always Representative for Non-Normal Distributions**: If the data is not normally distributed (i.e., if it's heavily skewed or has multiple peaks), the mean may not represent the dataset effectively. In such cases, the median or mode might be a better measure.
4. **Does Not Handle Categorical Data**: The mean cannot be used for categorical data (like favourite colours or types of animals), as it requires numerical values to perform the calculation.

Formulas of Mean:

1. Ungrouped Data: Direct method: $\boldsymbol{\overline{x} = \frac{\sum x}{n}}$

Short cut method: $\boldsymbol{\overline{x} = \frac{\sum di}{n}}$

2. Grouped data: Direct method: $\boldsymbol{\overline{x} = \frac{\sum fixi}{n}}$

Short -cut method: Mean = $\mathbf{A} + \frac{\sum fidi}{n} * C$

Example no.1: Find the average mean from the following 20,25,18,25,10,15,25.

Solution:

$\sum x = 20 + 25 + 18 + 25 + 10 + 15 + 25.$

$\bar{x} = \frac{\sum x}{n}$

$= 138 / 7$

=19.71

Example no.2: The following is a distribution of number of Students is 50 families. Find average.

Students	**0**	**1**	**2**	**3**	**4**	**5**
Frequency	**2**	**5**	**10**	**15**	**10**	**15**

Solution:

Students	**Frequency**	**fixi**
0	**2**	**0**
1	**5**	**5**
2	**10**	**20**
3	**15**	**45**
4	**10**	**40**
5	**15**	**75**
Total	**n = 57**	$\sum fixi$ **= 185**

$$\overline{x} = \frac{\sum fixi}{n}$$

$$\overline{x} = \frac{185}{57}$$

$$\overline{x} = 3.246$$

Example no.3: the frequency distribution of the marks in economics of 100 students is given below, Obtain average marks from it.

Marks	0-10	10-20	20-30	30-40	40-50
Freq.	5	10	15	10	5

Solution:

Marks	Freq	M.V (x)	fixi
0-10	5	5	25
10-20	10	15	150
20-30	15	25	375
30-40	10	35	350
40-50	5	45	225
Total	n = 45		$\sum fixi$ =1125

$$\overline{x} = \frac{\sum fixi}{n}$$

$$\overline{x} = \frac{1125}{45}$$

$$\overline{x} = 25$$

Example no. 4: find the Mean from the following:

Class	0-20	20-40	40-60	60-80	80-100
Freq	10	15	20	25	30

Solution:

Class	Freq	M.V (x)	di= $\left(\frac{x-A}{c}\right)$
			A =25
			C= 5
				fidi

0-20	10	10	-3	-30
20-40	18	15	-1	-18
40-60	22	25	0	0
60-80	28	35	1	28
80-100	34	45	3	102
Total	n = 112			$\sum fidi$ =82

$$\textbf{Mean} = \mathbf{A} + \frac{\sum fidi}{n} * C$$

$$= 25 + \frac{82}{112} * 5$$

$$= 25 + 3.66$$

$$= 28.66$$

4.4 Median

The **median** is a type of measure of central tendency that tells us the "middle" value in a dataset when the values are arranged in ascending or descending order. It is a great measure to use when you want to avoid being misled by extreme values (also known as outliers) that can distort the average.

Here is how to understand and calculate the median:

1. **Arrange the Data**: Start by organizing the data points from the smallest to the largest (or vice versa).
2. **Find the Middle**:
   - If the number of data points is **odd**, the median is simply the value in the middle. For example, in the dataset 3, 7, 9, 11, 13, the median is 9 because it is the third value (the middle one) in this list of five numbers.
   - If the number of data points is **even**, the median is the average of the two middle values. For instance, in the dataset 4, 7, 10, 13, the two middle numbers are 7 and 10. The median would be the average of these two values: (7 + 10) / 2 = 8.5.

Why use the Median?

- The median is especially useful when there are **outliers** or **extreme values** in the data. For example, if one student scored 100% on the test while everyone else scored much lower, the median will still reflect the "typical" score better than the mean (average).
- It is easy to calculate, even with large datasets, and it gives us a clear representation of where the centre of the data lies.

In summary, the **median** is a reliable way to determine the middle of a data set and works well for data that may be uneven or has outliers.

4.4.1. Merits and Demerits of the Median

Merits of the Median:

1. **Not Affected by Outliers**: One of the biggest advantages of the median is that it is **not influenced by extreme values** (outliers). For example, in a dataset where most values are close together, but one value is extremely high or low, the median remains unaffected, providing a more accurate representation of the "typical" value.

2. **Useful for Skewed Distributions**: The median is particularly useful when dealing with **skewed data**. In datasets where the values are not evenly distributed, the median gives a better idea of the central value compared to the mean, which might be distorted by outliers or extreme values.

3. **Easy to Understand**: The median is simple to understand, as it represents the middle value. This makes it a useful measure when you want to communicate the central tendency to a broad audience, especially when dealing with large datasets.

4. **Works with Ordinal Data**: The median can be used for **ordinal data** (data that has a meaningful order, but the distances between values are not consistent), unlike the mean, which requires numerical values with consistent interval.

Demerits of the Median:

1. **Ignores All Data Points Except the Middle**: The biggest drawback of the median is that it only considers the middle values in a dataset and **ignores the rest**. This means that it does not fully represent the entire dataset. If there are multiple values far from the middle, they are not taken into account, which could be a limitation in some cases.

2. **Less Precise in Symmetric Distributions**: In datasets that are **symmetrical** (i.e., bell-shaped or normally distributed), the median is less precise than the mean. The mean would give a more accurate measure of central tendency since it takes into account all data points, unlike the median, which might not reflect the distribution as well.

3. **Can Be Difficult to Calculate with Even Numbers**: When there is an even number of values, calculating the median requires averaging the two middle values, which can be a bit more complex than finding the mode or mean.

4. **Does not Represent the Data as Fully as the Mean**: Since the median only focuses on the middle value(s), it might not always reflect the spread or variability of the data, especially if the dataset contains values that are far from the centre.

Formula of median:

For ungrouped data: $\mathbf{M} = \left(\frac{n+1}{2}\right)^{\text{th}}$ **Observation.**

For grouped data : $\mathbf{M} = \left(\frac{n}{2}\right)^{\text{th}}$ **Observation**

$$\mathbf{M} = \mathbf{L} + \frac{\frac{n}{2} - cfi}{f} * \mathbf{C}$$

Example no.5: Find the median from the following: 50,40,25,36,24,35,13,62,.

Solution:

Arranging the observation in ascending order.

13,24,25,35,36,50,62.

n = 7

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

= $\left(\frac{7+1}{2}\right)$th Observation

= 4 th observation

M = 35

Example No. 6. Find the median from the following: 3,8,9,3,5,4,10,11.

Arranging the observation in ascending order.

3,3,4,5,8,9,10,11.

n = 8

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

= $\left(\frac{8+1}{2}\right)$th Observation

= 4.5th observation

= 4th + 5th / 2

= 5 + 8 /2

M = 13/2

M = 6.5

Example no. 7 Find the median of the distribution.

X	0	1	2	3	4	5	6	7	8
F	25	33	27	40	17	13	8	6	6

Solution:

X	F	cfi
0	25	25
1	33	58
2	27	85
3	40	125
40	17	142
5	13	155
6	8	163
7	6	169
8	6	175
Total	n = 175	

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

$$= \left(\frac{175+1}{2}\right)^{\text{th}} \text{ Observation}$$

= 88 th observation

= 3

Example no. 8: find the median from the following distribution:

Class	0-10	10-20	20-30	30-40	40-50	Total
Freq	10	20	40	30	50	150

Class	**Frequency**	**Cfi**
0-10	10	10
10-20	20	30
20-30	40	70
30-40	30	100
40-50	50	150
Total	**n = 150**	

$$M = \left(\frac{n}{2}\right)^{\text{th}} \text{ Observation.}$$

$$= \left(\frac{150}{2}\right)^{\text{th}} \text{ Observation}$$

= 75th observation

From the column 75th observation lies in the class 30-40. Taking limit points median class = 30-40.

$$M = L + \frac{\frac{n}{2} - cfi}{f} * C$$

L = 30, f = 3, cfi= 70

$$= L + \frac{\frac{n}{2} - cfi}{f} * C$$

$$= 30 + \frac{\frac{150}{2} - 70}{30} * 10$$

= 31.67

4.5 Mode

The **mode** is the measure of central tendency that identifies the most **frequent** value in a dataset. In simple terms, it's the number that appears most often. Unlike the mean and median, which are based on the values themselves, the mode is concerned with how often a particular value occurs.

How to Find the Mode:

1. **List the Data**: Write down all the values in the dataset.
2. **Count the Frequency**: Identify how many times each value appears.
3. **Identify the Most Frequent Value**: The value that appears the most is the mode.

For example, in this dataset:
2,4,4,6,7,7,7,82, 4, 4, 6, 7, 7, 7, 82,4,4,6,7,7,7,8

- 4 appears **twice**
- 7 appears **three times**
- 2, 6, and 8 each appear once.

In this case, the **mode** is 7, because it appears more frequently than any other number.

Types of Mode:

- **Unimodal**: A dataset with one mode (e.g., 4, 4, 5, 6, 7 – mode is 4).
- **Bimodal**: A dataset with two modes (e.g., 3, 3, 5, 5, 7 – modes are 3 and 5).
- **Multimodal**: A dataset with more than two modes (e.g., 2, 4, 4, 5, 5, 7 – modes are 4 and 5).
- **No Mode**: If all values appear only once, the dataset has no mode (e.g., 1, 2, 3, 4, 5).

Why use the Mode?

- The **mode** is particularly useful when you're dealing with **categorical data** or when you want to know the most common item. For example, in a survey where people select their favourite colour, the mode will tell you which colour was chosen the most.
- It is also helpful when the data contains **outliers** that may affect the mean. The mode is not influenced by extreme values, making it a good choice for finding the most common value.

4.5.1. Merits and Demerits of the Mode

Merits of the Mode:

1. **Simple to Understand and Calculate**: The **mode** is the easiest measure of central tendency to calculate. It simply identifies the most frequent value in a dataset, which makes it easy to understand, especially for beginners.
2. **Applicable to All Types of Data**: Unlike the mean or median, which require numerical data, the mode can be used with **nominal** (categorical) data, such as

colours, brands, or types of animals. It's particularly useful when dealing with data where the goal is to know the most common category.

3. **Not Affected by Extreme Values**: The mode is **not influenced by outliers** or extreme values. Even if there are very high or low values in the dataset, the mode will only focus on the most frequent value, making it useful in certain situations where outliers might distort other measures like the mean.
4. **Helps Identify Popular Trends or Preferences**: The mode is ideal for finding the most popular item in a set. For example, in a survey asking people about their favourite ice cream flavour, the mode will show the flavour chosen by the most people.

Demerits of the Mode:

1. **May Not Be Unique**: The **mode** is not always unique. A dataset can have **multiple modes** (bimodal or multimodal) or no mode at all if all values occur with the same frequency. This can make interpretation more difficult in some cases.
2. **Does Not Reflect the Central Value Accurately**: The mode does not necessarily represent the "typical" value of a dataset, especially if the values are spread out. For example, in a dataset of test scores like 40, 50, 50, 70, 100, the mode is 50, but 50 is not representative of the overall performance of the class.
3. **Less Useful for Continuous Data**: The mode is not as useful for **continuous data** (e.g., heights, weights, or time), because it only works when values are repeated. Continuous data rarely has repeating values, so the mode might not give meaningful insights in such cases.
4. **Does Not Account for All Data Points**: Like the median, the mode only focuses on the most frequent value and ignores the rest of the dataset. This means that it doesn't fully capture the distribution or variability of the data.

Formula of mode for grouped data: $\mathbf{Mo} = \mathbf{L} + \frac{fm-f1}{2fm-f1-f2} * C$

Example no.9: Find mode of the following observation: 2,3,5,4,5,2,8,9,7,2,1.2.

Solution: Here 2 is repeated for maximum number.

Example no. 10: Find mode of the following observation

Observation	10	11	12	13	14	15	16	17
Frequency	9	15	24	20	11	13	40	35

In the given frequency distribution maximum frequency is 40 and its observation is 16.

Example no. 11: find mode of the following distribution

Class	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**
freq	**6**	**10**	**20**	**25**	**15**	**5**

Here, the maximum frequency is in 30-40 class.

$$Mo = L + \frac{fm-f1}{2fm-f1-f2} * C$$

$$Mo = 30 + \frac{25-20}{2(25)-20-15} * 10$$

$$= 30 + \frac{50}{15}$$

$$= 30 + 3.33$$

$$= \mathbf{33.33}$$

- **Exercise**

A. Answer the following question

1. Define measures of average, median and mode.

2. Give the characteristics of Good Average.

3. Give the merit and demerit of Mean?

4. Give the merit and demerit of Median?

5. Give the merit and demerit of Mode**?**

B. Sove the following question:

1. Find mean, median and mode from the following data: 10, 12, 09, 13, 15, 22, 28, 15.

2. Find mean, median and mode from the following data:

x	10	11	12	13	14	15	16
f	18	20	12	20	25	7	6

3. Find mean, median and mode from the following data:

Class	0-9	10-19	20-29	30-39	40-49	50-59
Freq	5	12	28	20	4	3

BBA
SEMESTER-2
BUSINESS STATISTICS
BLOCK: 2

Authors' Name: Dr. Dipak Sanki, Assistant Professor, KBS College, Vapi
Dr.Maulik Desai, Assistant Professor, K.S. School of Business Management,Ahmedabad.

Review (Subject): Dr.Vaibhav Shah, Professor, New L J Commerce College, Ahmedabad
Dr.Poonam Arora , Professor, SKIPS School of Business, Ahmedabad

Review (Language): Dr. Bhavna Trivedi , Assistant Professor, Dr. BAOU, Ahmedabad

Editor's Name: Prof. (Dr.) Manoj Shah
Professor and Director,
School of Commerce and Management,
Dr. Babasaheb Ambedkar Open University,
Ahmedabad.

Publisher's Name: Dr. Ajaysinh Jadeja
Registrar,
Dr. Babasaheb Ambedkar Open University
'JyotirmayParisar',Opp.ShriBalaji
Temple,Chharodi,Ahmedabad,382481,
Gujarat,India.

Edition: 2024 (First Edition)

ISBN: 978-93-5598-707-5

978-93-5598-707-5

UNIT-5 MEASURES OF DISPERSION

5.1 Introduction

Imagine you and your friends took a math test, and everyone got different scores. While some scored very high, others scored average or low. Now, what if you were asked, *"How spread out are these scores?"* This is where the concept of *dispersion* comes into play. Dispersion helps us understand how data is distributed around a central value, such as the mean or median.

Dispersion answers questions like:

- Are the values clustered closely together, or are they scattered far apart?
- How much variation exists in a dataset?
- Are there extreme values that deviate significantly from the rest?

In real life, dispersion is everywhere. From analysing income inequality in a country to understanding how consistent a cricket player's performance is, the measure of dispersion gives valuable insights into data variability.

Before diving deeper, let's warm up with an example:

Quick Activity:
Imagine two scenarios:

1. In a classroom, students scored 50, 51, 49, 52, and 50 on a test.
2. In another classroom, students scored 20, 80, 15, 95, and 50.

Though both have the same average (50), the variation in scores is strikingly different. In Scenario 1, the scores are close to each other. In Scenario 2, they are widely scattered. This variation is what we aim to measure using the tools of dispersion.

5.2 Characteristics of Dispersion

Dispersion refers to the extent to which data values in a dataset deviate or spread out from a central point (mean, median, or mode). It is a critical concept in statistics that helps us understand the variability and consistency of data. Let us explore the key characteristics of dispersion:

1. Measures Variability: Dispersion quantifies the extent of variation in a dataset. If the data points are closely clustered around the central value, the dispersion is low; if they are widely spread, the dispersion is high.

2. Complementary to Central Tendency: While measures of central tendency (mean, median, mode) describe the "average" behaviour of a dataset, dispersion provides insights into the "spread" or "range" of the data. Both are necessary for a comprehensive data analysis.
3. Non-Negative Values: Dispersion measures are always non-negative. A dispersion value of zero indicates no variability (all data points are identical).
4. Sensitivity to Outliers: Some measures of dispersion, such as range and standard deviation, are highly influenced by extreme values (outliers). Others, like the interquartile range (IQR), are less affected.
5. Units of Measurement: Dispersion is often expressed in the same units as the original data. For example, if data is measured in kilograms, measures like standard deviation will also be in kilograms.
6. Summarizes Data Spread: Dispersion provides a single value to describe the degree of variation in the dataset, making it easier to compare datasets.
7. Forms the Basis for Advanced Analysis: Many advanced statistical techniques, such as hypothesis testing and regression analysis, rely on an understanding of data dispersion.
8. Types of Dispersion:
   - Absolute Measures: Provide the spread of data in absolute terms (e.g., range, standard deviation).
   - Relative Measures: Compare the spread relative to the central value or size of the dataset (e.g., coefficient of variation).

5.3 Range

Definition:
The range is the simplest measure of dispersion. It is defined as the difference between the maximum and minimum values in a dataset.

Range=Maximum Value−Minimum Value

Key Features of Range

1. **Ease of Calculation:**
   - The range is straightforward and requires only the highest and lowest values.
   - It gives a quick snapshot of data variability.
2. **Depends on Extreme Values:**
   - Since the range is based only on the two extreme values, it can be significantly influenced by outliers.
3. **Units of Measurement:**
   - The range is expressed in the same units as the original data.

4. **Limited Insight:**
   - The range does not provide information about the distribution of values between the extremes.

How to Calculate the Range

1. Identify the maximum value in the dataset.
2. Identify the minimum value in the dataset.
3. Subtract the minimum value from the maximum value.

Merits of Range

1. The range is easy to understand and calculate, requiring only two values: the maximum and minimum.
2. It provides a rapid assessment of the data's variability, making it suitable for initial data analysis.
3. The range is helpful when comparing the spread of two or more datasets.
4. It works well for small datasets where there aren't many values to analyse.
5. The range is expressed in the same units as the original data, making it easy to interpret.

Demerits of Range

1. A single extreme value (outlier) can drastically change the range, making it an unreliable measure in datasets with outliers.
2. The range only considers the maximum and minimum values, ignoring the distribution of the remaining data points.
3. It does not provide information about how the data is spread around the central tendency or whether the data points are evenly distributed.
4. In large datasets, the range loses its effectiveness as a measure of dispersion due to the potential for extreme values.
5. The range cannot be used to compare datasets with different units or scales without normalization.

Formula

R = $x_H - x_L$

Relative range = $\frac{xH - xL}{xH + xL}$

Example no. 1: Find range and relative range of prices.

150, 155, 160, 170, 165.

Solution:

$x_H = 170, x_L = 150$

R = x_H - x_L

= 170 - 150

= 20

Relative range = $\frac{\mathbf{xH - xL}}{\mathbf{xH + xL}}$

$= \frac{170 - 150}{170 + 150}$

$= \frac{20}{220}$

= 0.09

Example no. 2: Find the Range and Relative range from the following:

Observation	50	60	70	80	90	100	110
Frequency	5	10	18	22	25	24	25

Solution:

x_H = 170, x_L = 150

R = x_H - x_L

= 110 - 50

= 60

Relative range = $\frac{\mathbf{xH - xL}}{\mathbf{xH + xL}}$

$= \frac{110 - 50}{110 + 50}$

$= \frac{60}{160}$

= 0.375

Example no.3: Find the Range and Relative Range from the following:

Class	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**
freq	**5**	**14**	**19**	**25**	**23**	**30**

Solution:

x_H = 170, x_L = 150

R = x_H - x_L

= 90 - 30

= 60

$$\textbf{Relative range} = \frac{\mathbf{xH - xL}}{\mathbf{xH + xL}}$$

$$= \frac{90 - 30}{90 - 30}$$

$$= \frac{60}{120}$$

$$= 0.5$$

5.4 Quartile Deviation

Quartile deviation, also known as the semi-interquartile range, is a statistical measure that provides an indication of the spread or dispersion of the middle 50% of a dataset. It is particularly useful when dealing with data that may have outliers or skewed distributions, as it focuses on the central portion of the data and ignores extreme values.

In statistical analysis, understanding the variability of data is just as important as knowing the average or central value. Measures like the mean or median provide information about the central tendency, but they do not reveal how spread out or clustered the data points are. Quartile deviation helps address this by focusing on the interquartile range (IQR), which is the range between the first quartile (Q1) and third quartile (Q3).

- First Quartile (Q1): The value below which 25% of the data points lie.
- Third Quartile (Q3): The value below which 75% of the data points lie.

The quartile deviation is simply half the interquartile range, making it a concise measure of the variability in the central portion of the dataset.

Since quartile deviation considers only the middle 50% of the data, it is less sensitive to outliers and skewed data compared to other measures of spread, such as range or standard deviation. This makes it especially valuable while analysing datasets that are not symmetrically distributed or that contain extreme values that might otherwise distort the results.

Merits of Quartile Deviation:

1. Quartile deviation is resistant to the effects of outliers.
2. It focuses on the spread of the middle 50% of the data, making it more reliable for skewed distributions.
3. The calculation of quartile deviation is simple and easy to understand.
4. It provides a clear and intuitive measure of central data variability.
5. It is useful in scenarios where extreme values are not relevant to the analysis.

Demerits of Quartile Deviation:

1. It ignores extreme values, which may be important in some analyses.

2. Quartile deviation provides limited information about the total spread of the data.
3. It does not reflect the symmetry or skewness of the distribution.
4. It may not be sensitive or informative for uniformly distributed data.
5. The quartile deviation may be less reliable for very small datasets.

Formula of Qd

Quartile deviation (Qd) = $\frac{Q3-Q1}{2}$

Co-efficient of quartile deviation = $\frac{Q3-Q1}{Q3+Q1}$

Example no. 4: Find quartile deviation and coefficient of quartile deviation: 30, 10, 20, 35, 45, 55 and 72.

Arranging the observation 10,20,30,35,45,55,72.

$Q_1 = \left(\frac{n+1}{4}\right) th\ observation$

$= \left(\frac{7+1}{4}\right) th\ observation$

= 2 th observation

= 20

$Q3 = 3\left(\frac{n+1}{4}\right) th\ observation$

$= 3\left(\frac{7+1}{4}\right) th\ observation$

= 3 (20)

= 60

Quartile deviation (Qd) = $\frac{Q3-Q1}{2}$

$= \frac{60-20}{2}$

= 20

Example no.5: Find quartile deviation and coefficient of quartile deviation:

x	5	6	8	9	8	7	6	5	4
f	**20**	**40**	**60**	**40**	**50**	**35**	**25**	**15**	**10**

Solution:

X	F	Cfi
5	**20**	**20**
6	**40**	**60**

8	**60**	**120**
9	**40**	**160**
8	**50**	**210**
7	**35**	**245**
6	**25**	**265**
5	**15**	**280**
4	**10**	**290**
Total	**n= 290**	

$Q_1 = \left(\frac{n+1}{4}\right) th\ observation$

$= \left(\frac{290+1}{4}\right) th\ observation$

= 72.75th observation

= 8

$Q3 = 3\left(\frac{n+1}{4}\right) th\ observation$

$= 3\left(\frac{290+1}{4}\right) th\ observation$

= 3 (72.75)

= 218.25

= 245

Quartile deviation (Qd) = $\boldsymbol{\frac{Q3-Q1}{2}}$

$\boldsymbol{= \frac{245-8}{2}}$

= 118.5

Example no.5: Find quartile deviation and coefficient of quartile deviation:

Class	Frequency	Cumulative Frequency
5-10	15	15
10-15	32	47
15-20	21	68
20-25	16	84
25-30	25	109
30-35	36	145

$Q_1 = \left(\frac{n}{4}\right) th\ observation$

$= \left(\frac{145}{4}\right) th\ observation$

= 36.25 th observation

= class (10 – 15)

$$Q1 = L + \frac{\frac{n}{2} - cfi}{f} * C$$

L = 10, f = 32, cfi= 15

$$= L + \frac{\frac{n}{2} - cfi}{f} * C$$

$$= 10 + \frac{\frac{145}{2} - 15}{32} * 5$$

= 13.32

$$Q3 = 3\left(\frac{n+1}{4}\right) th\ observation$$

$$= 3\left(\frac{145}{4}\right) th\ observation$$

= 3 (36.25)

= 108.75

= class (25-30)

$$Q3 = L + \frac{\frac{3n}{2} - cfi}{f} * C$$

L = 25, f = 32, cfi= 15

$$= 10 + \frac{\frac{3(145)}{2} - 104}{25} * 5$$

= 25.95

Quartile deviation (Qd) = $\boldsymbol{\frac{Q3 - Q1}{2}}$

$$= \boldsymbol{\frac{25.95 - 13.32}{2}}$$

= 6.315

Coefficient of QD = QD / (Q3 + Q1)

= 6.315 / (25.95 + 13.32)

= 0.161

❖ **Exercise:**

A. Answer the following questions:

1. What is measures of dispersion.

2. Write short note on characteristic of good measures of dispersion.

3. What is range.

4. Write merit and demerit of range.

5. What is quartile deviation.

6. Write merit and demerit of quartile deviation.

B. Solve the sum

Exercise on Range and Quartile Deviation (QD)

1. Calculate the Range and Quartile Deviation for the following dataset:

25,30,35,40,45,50,55,60,65,7025, 30, 35, 40, 45, 50, 55, 60, 65, 7025,30,35,40,45,50,55,60,65,70

2. The heights (in cm) of 10 people are as follows:

160,165,170,172,175,180,185,190,192,200160, 165, 170, 172, 175, 180, 185, 190, 192, 200160,165,170,172,175,180,185,190,192,200

- Calculate the Range of this dataset.
- Calculate the Quartile Deviation.

3. Given the following test scores:

50,55,60,65,70,75,80,85,90,9550, 55, 60, 65, 70, 75, 80, 85, 90, 9550,55,60,65,70,75,80,85,90,95

- Find the Range.
- Find the Quartile Deviation.

4. For the following data on the number of books owned by 10 students:

4,8,10,12,14,16,18,20,22,244, 8, 10, 12, 14, 16, 18, 20, 22, 244,8,10,12,14,16,18,20,22,24

- Calculate the Range.
- Calculate the Quartile Deviation.

5. The ages of 12 participants in a survey are given as:

15,18,20,22,24,26,28,30,32,34,36,3815, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 3815,18,20,22,24,26,28,30,32,34,36,38

- Find the Range.
- Find the Quartile Deviation.

Final Answers:

1. Range: 45, Quartile Deviation: 12.5
2. Range: 40, Quartile Deviation: 10
3. Range: 45, Quartile Deviation: 12.5
4. Range: 20, Quartile Deviation: 5
5. Range: 23, Quartile Deviation: 7

C. **MCQs on Range and Quartile Deviation**

1. **What does the Range of a dataset represent?**
   - a) The difference between the highest and lowest values in the dataset.
   - b) The middle value of the dataset.
   - c) The sum of all data points divided by the number of data points.
   - d) The difference between the first and third quartiles.
2. **Which measure of dispersion is less affected by extreme values?**
   - a) Range
   - b) Quartile Deviation
   - c) Variance
   - d) Standard Deviation
3. **Which of the following is true about Quartile Deviation?**
   - a) It measures the spread of the entire dataset.
   - b) It is calculated using the first and third quartiles.
   - c) It is highly sensitive to extreme outliers.
   - d) It is used to find the average of the data points.
4. **When is Range considered a less reliable measure of dispersion?**
   - a) When the dataset is small.
   - b) When the data is highly skewed or contains outliers.
   - c) When the data has equal spread.
   - d) When the mean of the dataset is required.
5. **What does a high Quartile Deviation indicate about a dataset?**
   - a) The data is clustered around the central value.
   - b) There is a significant spread in the middle 50% of the data.

- c) The data has very few outliers.
- d) The data is very consistent.

6. **What is a major limitation of the Range?**
   - a) It uses only the two extreme values in the dataset.
   - b) It does not give an indication of the spread of the central data.
   - c) It is difficult to calculate.
   - d) It requires the data to be in ascending order.
7. **Which of the following is NOT a property of Quartile Deviation?**
   - a) It considers only the central 50% of the data.
   - b) It is not influenced by extreme values.
   - c) It is always larger than the range.
   - d) It can be used to describe the spread of data.
8. **The Range of a dataset is calculated by subtracting the:**
   - a) Mean from the median.
   - b) Maximum value from the minimum value.
   - c) First quartile from the third quartile.
   - d) Average of the data from the maximum value.
9. **If a dataset has a small Quartile Deviation, it means:**
   - a) The data points are widely spread.
   - b) The data is tightly packed in the middle.
   - c) There are significant outliers in the dataset.
   - d) The data follows a normal distribution.
10. **Which measure of dispersion is most commonly used to measure the consistency of data in the middle 50%?**
    - a) Range
    - b) Mean
    - c) Quartile Deviation
    - d) Standard Deviation

Final Answers:

1. a) The difference between the highest and lowest values in the dataset.
2. b) Quartile Deviation

3. b) It is calculated using the first and third quartiles.
4. b) When the data is highly skewed or contains outliers.
5. b) There is a significant spread in the middle 50% of the data.
6. a) It uses only the two extreme values in the dataset.
7. c) It is always larger than the range.
8. b) Maximum value from the minimum value.
9. b) The data is tightly packed in the middle.
10. c) Quartile Deviation

UNIT-6 STANDARD DEVIATION

6.1. Introduction

Standard deviation is a statistical measure that helps us understand the spread or dispersion of a data set. While the mean gives us a central value, it doesn't tell us how the individual data points are distributed around that central point. Standard deviation fills this gap by showing how much the data deviates, on average, from the mean.

In simple terms, if the data points are close to the mean, the standard deviation will be low, indicating little variability. On the other hand, if the data points are spread out over a wide range, the standard deviation will be high, reflecting greater variability. It is an essential concept in statistics, widely used in areas such as research, finance, and quality control, as it provides insight into the consistency or reliability of data.

How is Standard Deviation Calculated?

1. **Find the Mean**: Add up all the data points and divide by the total number of points.
2. **Calculate the Differences**: Subtract the mean from each data point to find how far each value is from the mean.
3. **Square the Differences**: Square each of the differences to eliminate negative numbers.
4. **Find the Average of Squared Differences**: This is called the variance.
5. **Take the Square Root of the Variance**: The final step is to take the square root of the variance to get the standard deviation.

6.2. Definition & Meaning

Here are a few definitions of **Standard Deviation** by notable statisticians:

1. **Karl Pearson**: "Standard deviation is the square root of the arithmetic mean of the squares of the deviations of the items from their arithmetic mean."
2. **Sir Ronald A. Fisher**: "The standard deviation is a measure of the scatter or spread of data, showing how much individual observations deviate from the mean."
3. **William S. Gosset (Student)**:"The standard deviation is a quantity that represents the degree of variation or dispersion of a set of data points."
4. **John Tukey**: "The standard deviation is a measure of the spread of a set of numbers; it is the square root of the average squared deviation from the mean."

These definitions focus on understanding standard deviation as a tool to measure how much data points deviate from the central value (mean) in a dataset, helping in assessing the variability or consistency of the data.

Meaning of Standard Deviation

Standard deviation is a statistical measure that tells us how spread out or dispersed the values in a data set are around the mean. It gives us an idea of the variability or consistency of the data. A **low standard deviation** means that the data points are close to the mean, indicating less variability, while a **high standard deviation** means that the data points are spread out over a wider range, indicating greater variability.

In simpler terms, it helps us understand whether the data is tightly clustered or widely scattered, providing insight into the predictability or consistency of the dataset.

Method of standard deviation

For ungrouped:

$$\mathbf{S} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$\mathbf{S} = \sqrt{\frac{\sum x^2}{n} - (\frac{\sum x}{n})^2}$$

$$\mathbf{S} = \sqrt{\frac{\sum d^2}{n} - (\frac{\sum d}{n})^2}$$

For grouped data: (discrete)

$$\mathbf{S} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$\mathbf{S} = \sqrt{\frac{\sum fixi^2}{n} - (\frac{\sum fixi}{n})^2}$$

$$\mathbf{S} = \sqrt{\frac{\sum fidi^2}{n} - (\frac{\sum fidi}{n})^2}$$

For grouped data: (continuous)

$$S = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$S = \sqrt{\frac{\sum fixi^2}{n} - (\frac{\sum fixi}{n})^2}$$

$$S = \sqrt{\frac{\sum fidi^2}{n} - (\frac{\sum fidi}{n})^2} * C$$

6.3. Objective of Standard Deviation

The primary objective of standard deviation is to measure the extent of variation or dispersion in a data set. It helps to:

1. **Understand Data Variability**: Standard deviation provides a clear picture of how much individual data points deviate from the mean, helping to assess the consistency or unpredictability of data.
2. **Compare Data Sets**: It allows for comparisons between different data sets. Even if two sets have the same mean, the one with the higher standard deviation shows more variability.
3. **Measure Consistency**: In fields like quality control, finance, and research, standard deviation is used to assess the reliability and consistency of processes, investments, or experimental results.
4. **Identify Risk or Uncertainty**: In finance, for example, standard deviation is used to measure the risk associated with an investment. A higher standard deviation indicates greater potential for fluctuation, hence greater risk.
5. **Help in Decision Making**: By understanding how spread out data points are, it aids in making informed decisions, especially when dealing with uncertainty and variability in data.

6.4. Advantages of Standard Deviation

1. **Quantifies Variability**: Standard deviation provides a clear numerical value that reflects how spread out or concentrated the data is around the mean. This helps to understand the variability within the data set.
2. **Useful for Data Comparison**: It enables easy comparison between different data sets. By comparing the standard deviations, you can determine which data set has more consistency or variability, even if their means are similar.
3. **Widely Applicable**: Standard deviation is a versatile tool used in various fields such as statistics, economics, finance, engineering, and research to assess the spread of data and make informed decisions.
4. **Helps Identify Outliers**: Large deviations from the mean can be easily spotted, helping to identify unusual data points (outliers) that may affect the analysis or interpretation.
5. **Foundation for Other Statistical Measures**: Standard deviation is foundational in many other statistical techniques, such as hypothesis testing, regression analysis, and analysis of variance (ANOVA). It plays a critical role in understanding data distributions, such as normal distributions.
6. **Reflects Real-world Data**: Since many natural and social phenomena follow a bell-shaped curve (normal distribution), standard deviation provides an effective measure for understanding real-world data, such as exam scores, market prices, or production quality.
7. **Helps in Risk Assessment**: In finance and investment, standard deviation is used to measure the volatility or risk of an asset. A higher standard deviation suggests higher risk, helping investors make better decisions.

8. **Improves Predictive Accuracy**: By understanding the spread of data, standard deviation helps in building more accurate models and predictions, especially when data consistency is crucial for forecasting.

6.5. Disadvantages of Standard Deviation

1. **Sensitive to Outliers**: Standard deviation is heavily influenced by extreme values or outliers in the data. A single very large or very small value can significantly increase the standard deviation, making it less representative of the majority of the data points.
2. **Not Suitable for Non-Normal Distributions**: If the data does not follow a normal (bell-shaped) distribution, standard deviation may not be an accurate measure of variability. In such cases, other measures like the interquartile range or median absolute deviation might be more appropriate.
3. **Complex Calculation**: The process of calculating standard deviation can be mathematically intensive and challenging for larger datasets, especially if they involve complex operations like squaring differences and square rooting. This can make it difficult for beginners to understand and apply without sufficient knowledge of statistics.
4. **Doesn’t Account for Skewness**: Standard deviation doesn’t provide information about the shape of the distribution (skewness). A highly skewed distribution might have a small mean but a larger standard deviation, which can be misleading.
5. **Sensitive to Small Sample Sizes**: When working with smaller sample sizes, standard deviation may not be representative of the population as a whole. This can introduce bias, making it less reliable for generalizations.
6. **Not Robust Against Variability Types**: If the variability within the dataset is caused by factors that are not linear (e.g., cyclical or seasonal), standard deviation might not fully capture the underlying patterns or relationships.
7. **Requires Assumptions About Data**: The computation of standard deviation relies on the assumption that the data is normally distributed. If this assumption doesn’t hold (for example, in highly skewed or multimodal data), the standard deviation may not be a suitable measure of dispersion.
8. **Doesn’t Reflect Relative Values**: While standard deviation provides a measure of spread, it does not inherently provide a relative or comparative perspective between two or more datasets unless those datasets share similar means or are scaled appropriately.

6.6. Co-efficient of Variation (CV)

The **Co-efficient of Variation (CV)** is a statistical measure that expresses the **relative variability** of a data set in relation to its mean. It is useful when comparing the degree of variation between different datasets, especially when they have different units or vastly different means.

Co-efficient of variance = $\frac{s}{\bar{x}}$

Meaning: The Coefficient of Variation is the ratio of the standard deviation to the mean, multiplied by 100 to express it as a percentage. It is used to measure the **relative dispersion** of data, which means it shows how large the standard deviation is compared to the mean. A higher CV indicates greater variability in relation to the mean, while a lower CV suggests less variability.

Key Points:

- The **CV is dimensionless** (it does not have units), which makes it particularly useful for comparing the variability of datasets with different units or scales.
- **Lower CV**: Indicates less variation relative to the mean (more consistency).
- **Higher CV**: Indicates greater variability relative to the mean (less consistency).

Uses of Co-efficient of Variation:

1. **Comparing Datasets**: CV is useful for comparing the variability of different datasets that have different units or means.
2. **Risk Assessment in Finance**: In finance, CV is used to compare the risk (volatility) of different investments, as it shows the risk per unit of return.
3. **Quality Control**: CV helps to assess consistency in manufacturing or other processes, especially when the mean values differ significantly.

Advantages:

- It provides a **relative measure** of variability.
- **Comparative tool** across datasets of different units or magnitudes.

Disadvantages:

- **Sensitive to small means**: If the mean is very small or close to zero, the CV can become very large or misleading.
- **Not applicable for negative data**: Since CV is based on the mean and standard deviation, it is not meaningful for datasets with negative values.

Example no. 1: The monthly expenses of students are given. Find Standard deviation and coefficient of S.D from the following

50,40,30,25,15,60,70,30.

Solution:

Expenses	**x - $\bar{x}$**	$(x-\bar{x})^2$
50	10	100
40	0	0
30	-10	100
25	-15	225
15	-25	625
60	20	400
70	30	900
30	-10	100
$\sum x$ = **320**		$\sum x-\bar{x}^2$ = **2450**

Mean = $\frac{\sum x}{n}$

$= \frac{320}{8}$

= 40

S = $\sqrt{\frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{n}}}$

= $\sqrt{\frac{\mathbf{2450}}{\mathbf{8-1}}}$

= 18.70

Example no. 2: The monthly expenses of students are given. Find Standard deviation and coefficient of S.D from the following

220,180,200,240,260,200,300,170,260.

Solution:

Expenses	**d = x - A**	**d ^{2}**
220	-6	36
180	-46	2116
200	-26	676
240	14	196
260	34	1156
200	-26	676
300	74	5476
170	-56	3136
260	34	1156
$\sum \boldsymbol{x}$ **= 2030**	$\sum \boldsymbol{d} = -\mathbf{4}$	**∑d ^{2} =14624**

Mean = $\frac{\sum \boldsymbol{x}}{\boldsymbol{n}}$

= $\frac{\mathbf{2030}}{\mathbf{9}}$

= 225.56

S = $\sqrt{\frac{\sum d^2}{n} - (\frac{\sum d}{n})^2}$

S = $\sqrt{\frac{14624}{9} - (\frac{-4}{9})^2}$

= 42.7525

Coefficient of variance = $\frac{s}{\bar{x}}$

$$= \frac{42.7525}{225.5556} * \ 100$$

= 18.95 %

Example no. 3: The following distribution shows the number of accidents in 30 cities during a week. Find standard deviation of numbers of accidents.

No of accidents	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number of cities	**2**	**3**	**4**	**7**	**5**	**5**	**3**	**1**

Solution:

x	**f**	**fixi**	**fixi^2**
1	2	2	2
2	3	6	12
3	4	12	36
4	7	28	112
5	5	25	125
6	5	30	180
7	3	21	147
8	1	8	64
Total	n = 30	**fixi** =132	**fixi^2** = 678

Mean = $\frac{\sum fixi}{n}$

$= \frac{132}{30}$

= 4.4

S = $\sqrt{\frac{\sum fixi^2}{n} - (\frac{\sum fixi}{n})^2}$

S = $\sqrt{\frac{678}{8} - (\frac{132}{8})^2}$

= 1.8308

Coefficient of variance = $\frac{s}{\bar{x}}$

$= \frac{1.8308}{4.4} * \mathbf{100}$

= 41.61 %

Example no. 4 : Find S.D of the following distribution

Class	Frequency
100-200	15
200-300	20
300-400	10
400-500	18
500-600	14
600-700	23
700-800	15
800-900	20
Total	135

Solution:

Class	**Frequency**	**M.V = x**	**d = x-A / c**	**fidi**	**fidi^2**
100-200	15	150	-4	-60	240
200-300	20	250	-3	-60	180
300-400	10	350	-2	-20	40
400-500	18	450	-1	-18	18
500-600	14	550	0	0	0
600-700	23	650	1	23	23
700-800	15	750	2	30	60
800-900	20	850	3	60	180
Total	**135**			**fidi = -45**	**fidi^2 = 741**

Mean = A + $\frac{\sum fidi}{n} * c$

= 550 + $\frac{-45}{135} * 100$

= 516.6667

S = $\sqrt{\frac{\sum fixi^2}{n} - (\frac{\sum fixi}{n})^2}$

S = $\sqrt{\frac{741}{135} - (\frac{-45}{135})^2}$

= 232.7641

Coefficient of variance = $\frac{s}{\bar{x}}$

= $\frac{232.7641}{516.6667} * \ 100$

= 45.05 %

❖ Exercise

A. Answer the following questions.

1. What is standard deviation.

2. Explain why standard deviation is more important in statistics.

3. Write the objectives of S.D.

4. Write advantage and disadvantage of S.D.

5. Write the formulas of S.D.

B. Sum on Standard Deviation

1. Calculate the standard deviation for the following data:

5, 7, 10, 12, 15, 18, 20

Answer: σ = 5.01

2. Find the standard deviation for the following frequency distribution:
Class Interval: 10-20, 20-30, 30-40, 40-50, 50-60
Frequency: 4, 6, 8, 10, 12

Answer: σ = 14.14

3. For the dataset with grouped intervals:
Class Midpoints: 5, 15, 25, 35, 45
Frequency: 2, 3, 5, 4, 6

Answer: σ = 13.23

4. Calculate the standard deviation for the dataset:
25, 30, 35, 40, 45, 50

Answer: σ = 7.91

5. Given the population dataset:
4, 8, 12, 16, 20
Find the standard deviation.

Answer: σ = 5.48

6. For the sample dataset:
10, 15, 20, 25, 30
Compute the standard deviation.

Answer: s = 8.37

7. If the variance of a dataset is 49, find the standard deviation.

Answer: σ = 7

8. A company measures the daily sales (in units) over a week:
50, 55, 60, 65, 70, 75, 80
Compute the standard deviation of daily sales.

Answer: σ = 10.61

9. Find the standard deviation for the frequency distribution:
Class Interval: 0-10, 10-20, 20-30, 30-40, 40-50
Frequency: 3, 5, 8, 6, 4

Answer: σ = 12.13

10. Two datasets have the following values:
Dataset A: 5, 10, 15, 20, 25
Dataset B: 10, 12, 14, 16, 18
Calculate the standard deviation for both datasets and compare.

Answer:
- Dataset A: σ = 7.07
- Dataset B: σ = 2.83

11. Calculate the C.V. if the mean of a dataset is 50 and the standard deviation is 10.
Answer: 20%

12. For a dataset, the mean is 25 and the standard deviation is 5. Find the C.V..

Answer: 20%

13. The heights (in cm) of 10 students are:

150,155,160,165,170,175,180,185,190,195. Calculate the **C.V.**.

Answer: 10.37%

14. The weights (in kg) of 8 parcels are: 12,15,18,21,24,27,30,33. Calculate the **standard deviation** and determine the **C.V.**.

Answer:

- **Standard Deviation**: 7.227.227.22
- **C.V.**: 28.88%

UNIT-7 MEAN DEVIATION

7.1. Introduction

When we analyse data, one important aspect we often look for is how spread out the values are around a central point like the mean or median. This concept of spread or variability is essential because it helps us understand the consistency of the data. For instance, if you're analysing test scores in a class, the spread will tell you whether students performed consistently or if there were significant differences in their scores.

In statistics, understanding how data behaves is crucial for making informed decisions. While measures like the mean and median help us identify the central tendency, they alone cannot give us a complete picture of a data set. We also need to understand the *dispersion* or variability—the extent to which data points differ from the central value. This is where measures like **Mean Deviation** come into play.

Mean Deviation is one of the simplest methods to measure dispersion. It tells us the average distance of all data points from a central value, such as the mean or median, without considering whether the data points are above or below this central value. By focusing on absolute differences, it provides a clear and unbiased picture of variability in the data.

This chapter introduces the concept of Mean Deviation, its calculation, and its significance in analysing data. We will explore its formula, steps to compute it, and practical applications, along with its advantages and limitations. By the end of the chapter, you will understand how Mean Deviation is used to assess the consistency and spread of data sets, helping you to make meaningful interpretations and comparisons.

7.2. Meaning and definition of Mean deviation

Definition of Mean Deviation:

Statisticians have provided formal definitions of Mean Deviation to emphasize its role in measuring dispersion. Some notable definitions include:

1. **A. L. Bowley**: *"The mean deviation is the average amount by which the items of a series deviate from the mean or median, ignoring the signs of deviation."*
2. **Spiegel (Murray R. Spiegel)**:*"Mean Deviation is the arithmetic mean of the absolute deviations of all the values in a dataset from a central value, usually the mean or median."*

3. **Prof. Boddington**: *"Mean deviation is the mean of the absolute differences between each value of a dataset and a measure of central tendency, like the mean or the median."*

These definitions highlight that Mean Deviation focuses on the absolute magnitude of deviations, making it an intuitive and straightforward measure of variability.

Meaning of Mean Deviation

Mean Deviation, also referred to as **Mean Absolute Deviation (MAD)**, is a statistical measure that calculates the average of the absolute differences between each data point in a dataset and a central value, such as the **mean** or **median**. It is used to assess how spread out or scattered the data points are around the central value.

In simple terms, it reflects the **average deviation** of data points from the chosen central point, ignoring whether the deviation is positive or negative. This approach provides a straightforward way to understand the consistency or variability within a dataset:

- **Smaller Mean Deviation**: Indicates that the data points are closely grouped around the central value, implying less variability.
- **Larger Mean Deviation**: Suggests that the data points are more spread out, showing greater variability.

Mean Deviation is commonly applied in fields such as quality control, risk assessment, and performance evaluation, where understanding the degree of variation is critical for decision-making.

7.3. Objectives of Mean Deviation

Mean Deviation serves several purposes in statistical analysis, helping to understand and interpret the variability within a dataset. The key objectives of Mean Deviation are:

1. **Measure Data Dispersion**: To quantify the spread or variability of data points around a central value, such as the mean or median.
2. **Simplify Deviation Analysis**: By using absolute values, it eliminates the issue of negative deviations cancelling out positive ones, making it easier to interpret the data.
3. **Evaluate Data Consistency**: To assess the uniformity or consistency of a dataset. A smaller Mean Deviation indicates high consistency, while a larger value suggests greater variability.
4. **Compare Variability**: To compare the dispersion of two or more datasets, especially when analysing different groups or categories.
5. **Support Decision-Making**: To provide insights into the degree of variation, which aids in making informed decisions in fields like finance, quality control, and risk management.
6. **Foundation for Further Analysis**: To serve as a basic tool for understanding data variability, forming the groundwork for more advanced statistical measures like variance and standard deviation.

By achieving these objectives, Mean Deviation becomes a valuable tool for statistical and practical data analysis in various disciplines.

7.4. Advantages and Disadvantages

Advantages of Mean Deviation

Mean Deviation offers several benefits as a measure of dispersion, making it a useful tool in statistical analysis. The key advantages are:

1. **Simplicity**: The calculation of Mean Deviation is straightforward and easy to understand, as it involves basic arithmetic and absolute differences.
2. **Eliminates Negative Deviations**: By using absolute values, it avoids the issue of negative deviations cancelling out positive ones, ensuring an accurate measure of variability.
3. **Focus on Actual Differences**: It directly reflects the average size of deviations, providing a clear understanding of how much data points vary from the central value.
4. **Applicable to Different Central Values**: Mean Deviation can be calculated around both the mean and the median, offering flexibility based on the nature of the data.
5. **Useful for Comparative Analysis**: It allows for an easy comparison of variability between two or more datasets, making it valuable for comparative studies.
6. **Practical in Real-Life Applications**: Mean Deviation is widely used in fields such as finance, quality control, and risk analysis, where understanding variability is critical.
7. **Effective for Small Data Sets**: For small datasets, Mean Deviation is a quick and efficient way to gauge dispersion without requiring complex computations.

These advantages make Mean Deviation a popular choice, especially when a simple and clear measure of data variability is required.

Disadvantages of Mean Deviation

While Mean Deviation is a useful and simple measure of dispersion, it also has some limitations:

1. **Less Sensitivity to Extreme Values**: Mean Deviation does not give as much weight to outliers or extreme values as other measures like **variance** or **standard deviation**, which can be a disadvantage in datasets where extreme values are significant.

2. **Not as Commonly Used as Other Measures**: In advanced statistical analysis, Mean Deviation is often considered less robust compared to variance or standard deviation. These measures are more widely used because they provide a deeper understanding of variability and are foundational for many statistical tests.

3. **Ignores the Direction of Deviation**: By using absolute values, Mean Deviation ignores whether the deviation is positive or negative. This can sometimes be a disadvantage when the direction of the deviation (e.g., above or below the mean) is important for analysis.

4. **Not Ideal for Skewed Data**: Mean Deviation can be less informative when the data is heavily skewed because it does not account for how data is distributed around the central value. In such cases, other measures like **standard deviation** might be more useful.
5. **Limited for Statistical Inference**: Mean Deviation is not suitable for statistical inference (such as hypothesis testing or regression analysis) because it lacks the mathematical properties that make variance and standard deviation useful in more complex analyses.

Despite these drawbacks, Mean Deviation is still a valuable tool for basic descriptive analysis and can be used effectively when a simple, easily interpretable measure of spread is needed.

7.5. Method of Measuring Mean Deviation

Mean deviation MD = $\frac{\sum|\boldsymbol{x}-\overline{\boldsymbol{x}}|}{\boldsymbol{n}}$

Where $[x - \bar{x}]$ = absolute deviation from the mean

n = number of observations.

The relative measure of dispersion obtained from mean deviation is called the coefficient of mean deviation and it is as follows.

Coefficient of mean deviation = $\frac{\mathbf{MD}}{\overline{\boldsymbol{x}}}$

Example no 1: Mean Deviation and coefficient of mean deviation from the following observation: 46,56,60,40,70,60,40,70,60,51,81.

X	$(\boldsymbol{X} - \overline{\boldsymbol{X}})$	$|\boldsymbol{x} - \overline{\boldsymbol{x}}|$
46	-12	12
56	-2	2
60	2	2
40	-18	18
70	12	12
60	2	2
51	-7	7
81	23	23
Total	**-**	**78**

Mean = $\frac{\sum \boldsymbol{x}}{\boldsymbol{n}}$

$= \frac{464}{8}$

= 58

Mean deviation MD = $\frac{\sum|\boldsymbol{x}-\overline{\boldsymbol{x}}|}{\boldsymbol{n}}$

$= \frac{78}{8}$

= 9.75

Coefficient of mean deviation = $\frac{\mathbf{MD}}{\overline{\boldsymbol{x}}}$

$= \frac{9.75}{58}$

= 0.1681

Example no. 2: Find mean deviation and coefficient of mean deviation for the following frequency distribution:

X	50	51	52	53	54	55	56	57
F	6	14	37	54	70	25	12	10

Solution:

X	**f**	**fixi**	$\|\boldsymbol{x}-\overline{\boldsymbol{x}}\|$	**f**$\|\boldsymbol{x}-\overline{\boldsymbol{x}}\|$
50	6	300	3.4956	20.9737
51	14	714	2.4956	34.9386
52	37	1924	1.4956	55.3377
53	54	2862	0.4956	26.7632
54	70	3780	0.5044	35.307
55	25	1375	1.5044	37.6096
56	12	672	2.5044	30.0523
57	10	570	3.5044	35.0439
Total	**n = 228**	$\sum \boldsymbol{fixi}$ **=12197**		$\sum \boldsymbol{f}\|\boldsymbol{x}-\overline{\boldsymbol{x}}\|$**=276.0263**

Mean = $\frac{\sum \boldsymbol{fixi}}{\boldsymbol{n}}$

$= \frac{\mathbf{12197}}{228}$

= 53.4956

Mean deviation MD = $\frac{\sum \boldsymbol{f|x-\bar{x}|}}{\boldsymbol{n}}$

$= \frac{276.0263}{228}$

= 1.2106

Coefficient of mean deviation = $\frac{\mathbf{MD}}{\boldsymbol{\bar{x}}}$

$= \frac{1.2106}{53.4956}$

= 0.0226

Example 3: Find the mean deviation for the following distribution:

Class Interval	**Frequency**
0-10	2
10-20	5
20-30	12
30-40	18
40-50	15
50-60	8
60-70	5

Solution:

Class	**f**	**X = mid-value**	**di** = $\frac{x-A}{c}$ **A = 35** **C= 10**	**Fidi**	$\|x - \bar{x}\|$	**f**$\|x - \bar{x}\|$
0-10	2	5	-3	-6	19.6	39.2
10-20	5	10	-2	-10	14.6	73
20-30	12	15	-1	-12	9.6	115.2
30-40	18	20	0	0	4.6	82.8
40-50	15	25	1	15	14.4	216

50-60	8	30	2	16	19.4	155.2
60-70	5	35	3	15	24.4	122
Total	**n = 75**	$\sum fixi$ **=12197**				$\sum f\lvert x-\bar{x}\rvert$**=900.4**

x̄ = A + Σfd/Σf * C

= 35 + -12/75 * 10

= 33.4

Mean deviation MD = $\frac{\sum f|x-\bar{x}|}{n}$

$$= \frac{900.4}{75}$$

= 12.0053

Coefficient of mean deviation = $\frac{\text{MD}}{\bar{x}}$

$$= \frac{12.0053}{33.4}$$

= 0.36

- **Exercise**

A. Answer the following questions

1. What is mean deviation?
2. Discuss the importance of mean deviation?
3. Write advantage of mean deviation?
4. Write disadvantage of mean deviation?
5. Write the objective of mean deviation?

B. Multiple Choice Questions (MCQs)

1. **What does Mean Deviation measure?**

   a) The total sum of deviations in a dataset

   b) The average distance of data points from a central value

   c) The square of the deviations from the mean

   d) The range of a dataset

2. **Which of the following is a central value used in calculating Mean Deviation?**

   a) Mean

   b) Median

   c) Mode

   d) Both a and b

3. **Why does Mean Deviation use absolute values of deviations?**

   a) To simplify the calculation process

   b) To ensure that negative deviations do not cancel out positive deviations

   c) To account for extreme values

   d) To give higher weight to central values

4. **Who defined Mean Deviation as "the average amount by which the items of a series deviate from the mean or median, ignoring the signs of deviation"?**

   a) Murray R. Spiegel

   b) A.L. Bowley

   c) Prof. Boddington

   d) Karl Pearson

5. **In Mean Deviation, a smaller value indicates:**

   a) Greater variability

   b) Less variability

   c) The presence of outliers

   d) High skewness

6. **Which of the following is a limitation of Mean Deviation?**

   a) It is difficult to calculate

   b) It ignores the direction of deviation

   c) It gives more weight to extreme values

   d) It is not useful for small datasets

7. **Mean Deviation can be applied in the following field(s):**

   a) Finance

   b) Quality control

   c) Risk assessment

   d) All of the above

8. **Which of the following advantages of Mean Deviation is true?**

   a) It eliminates negative deviations

   b) It gives more weight to extreme values

   c) It is ideal for skewed data

   d) It is suitable for statistical inference

9. **What does a larger Mean Deviation suggest?**

   a) Data points are closely grouped

   b) Data points are spread out

   c) Data has no variability

   d) Central value is incorrect

Answers

1. b) The average distance of data points from a central value
2. d) Both a and b
3. b) To ensure that negative deviations do not cancel out positive deviations
4. b) A.L. Bowley
5. b) Less variability
6. b) It ignores the direction of deviation
7. d) All of the above
8. a) It eliminates negative deviations
9. b) Data points are spread out

C. Short Numerical Problems

1. Calculate the Mean Deviation (MD) for the following data:

   X: {5,10,15,20,25}

2. Find the Coefficient of Mean Deviation if:

   Mean Deviation (MD) = 6.2

   Arithmetic Mean = 25

3. Given the following data, compute the Mean Deviation (MD):

   X: {12,14,18,22,24}

4. The Mean Deviation (MD) for a dataset is 4.8, and the Mean is 20. What is the Coefficient of Mean Deviation?

5. For the data X: {8,16,20,24,28} calculate the Mean Deviation (MD) and the Coefficient of Mean Deviation (using the average of X).

6. Solve the following questions:

Find mean deviation and coefficient of mean deviation for the following frequency distribution:

Class	0-9	10-19	20-29	30-39	40-49	50-59
Freq	11	23	46	70	40	20

Answer

1. **MD = 6**
2. **Coefficient of MD = 0.248**
3. **MD = 4**
4. **Coefficient of MD = 0.24**
5. **MD = 5.76, Coefficient of MD = 0.3**
6. **Mean Deviation (M.D.): 9.91**, **Coefficient of Mean Deviation: 35.39%**

UNIT-8 CORRELATION

8.1 Introduction

Correlation is a measure of the degree of relatedness of variables. It can help a business person determine, for example, whether the stocks of two companies rise and fall in any related manner. For a sample of pairs of data, correlation analysis can yield a numerical value that represents the degree of relationship of the two stock prices over time. In the transportation industry, is a correlation evident between the price of transportation and the weight of the object being shipped? If so, how strong are the correlations? In the area economics, how strong is the correlation between the price and demand of the commodity? In retail sales, are sales related to population density, number of competitors, size of the store, amount of advertising, or other variables?

The statistic ‘r’ is the Pearson product-moment correlation coefficient, named after the renowned scientist Karl Pearson (1857–1936), an English statistician who developed several coefficients of correlation along with other significant statistical concepts. The term ‘r’ is a measure of the linear correlation of two associated variables. It is a number that ranges from – 1 to + 1, representing the strength of the relationship between the variables. An ‘r’ value of + 1 denotes a perfect positive relationship between two sets of numbers. An ‘r’ value of – 1 denotes a perfect negative correlation, which indicates an inverse relationship between two variables. An ‘r’ value of 0 means no linear relationship is present between the two variables.

8.2 Objectives

Correlation is a statistical measure of finding relationships between two random variables or two sets of data. Correlation measures the strength of the linear relationship between two variables.

The key objectives of correlation include:

1. Determine the Nature and Strength of the Relationship
2. Identify Patterns and Trends
3. Support Prediction
4. Aid in Decision-Making
5. Explore Potential Causation (with Caution)
6. Evaluate Hypotheses

7. Measure the Interdependence of Variables
8. Inform Further Statistical Analysis

8.3 Definition

If there is cause and effect relationship between two variables and value of one variable is changed and due to that there is simultaneous change in the other variable, then these variables are said to be correlated with each other. The degree of relationship among them is called coefficient of correlation.

As we discussed coefficient of correlation is indicated by ‘r. The value of it always lies between – 1 to + 1.

Properties of Coefficient of Correlation

1. It is an absolute number, and it is free from any measurement unit.
2. The value of it always lies between – 1 to + 1.
3. The value of r^2 always lies between 0 and 1.
4. The Coefficient of Correlation is independent of change of origin and change of scale.

8.4 Types of Correlation

Correlation mainly divided as follows:

- **Positive Correlation**

  If a change among the variables happened in the same direction, then this type of correlation is called positive correlation.

  For example: Price and Supply of the commodity, Age of married couple, Height and Weight of growing child etc.

- **Negative Correlation**

  If a change among the variables happened in reverse direction, then this type of correlation is called negative correlation.

  For example: Price and Demand of the commodity, Volume and Pressure, Expenditure and Saving etc.

- **Lack of Correlation or No Correlation**

  If change among the variables happened in random manner, then this type of correlation is called lack of correlation or no correlation.

  For example: Increase Price of petrol and usage of petrol, Increase price of salt and usage of salt etc.

8.5 Uses of Correlation

The study of correlation is useful as follows:

1) The correlation coefficient helps in measuring the extent of the relationship between two variables in one figure.

2) Correlation analysis facilitates the understanding of economic behaviour and helps in locating the critically important variables on which others depend.

3) When two variables are correlated, the value of one variable can be estimated, given the value of another. This is done with the help of regression equations.

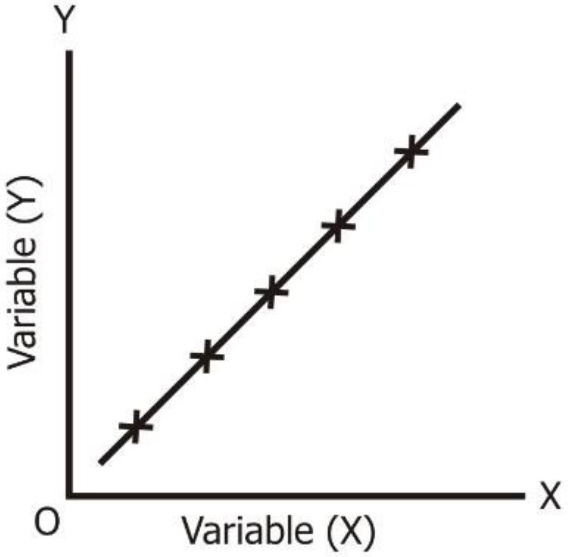
4) Correlation facilitates the decision-making in the business world. It reduces the range of uncertainty as predictions based on correlation are likely to be more reliable and near to reality.

8.6 Techniques of Measuring correlation

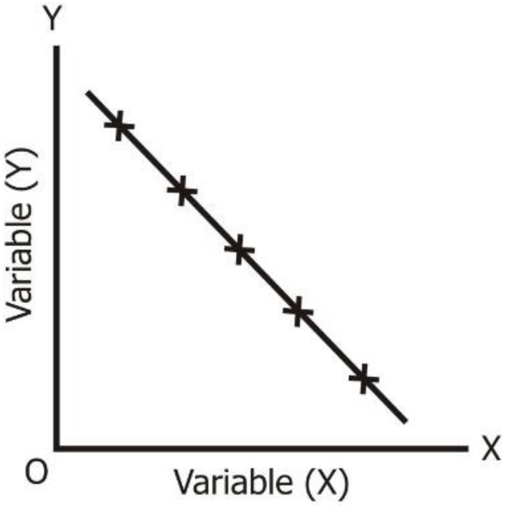
1) Scatter Diagram Method

In this method both variables are plotted on graph paper by taking the variables on X – Axis and Y – Axis respectively. After that by evaluating the nature of the graph one can decide the nature of the relationship among the variables as follows:

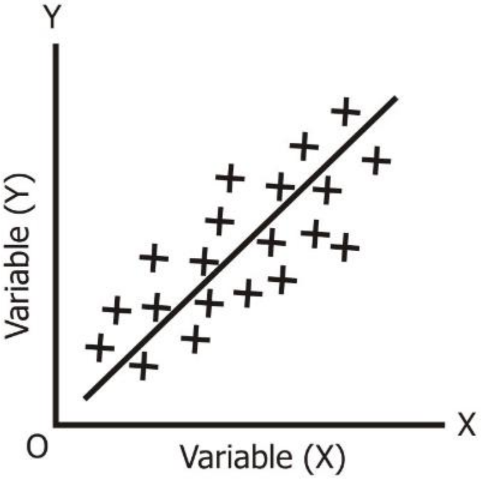
Perfect Positive Correlation



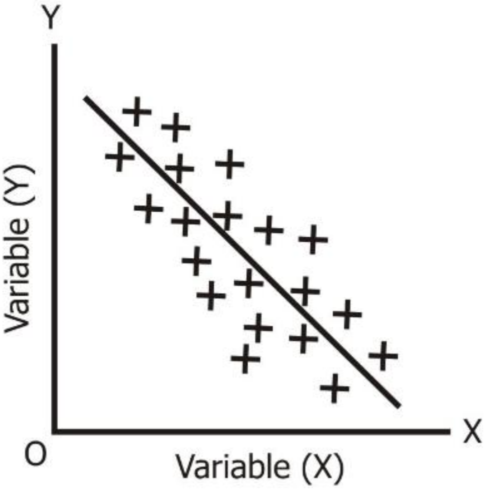
Perfect Negative Correlation



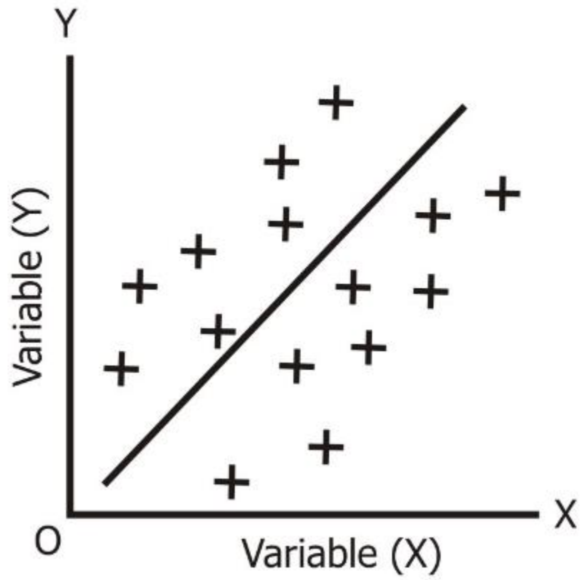
High Degree of Positive Correlation



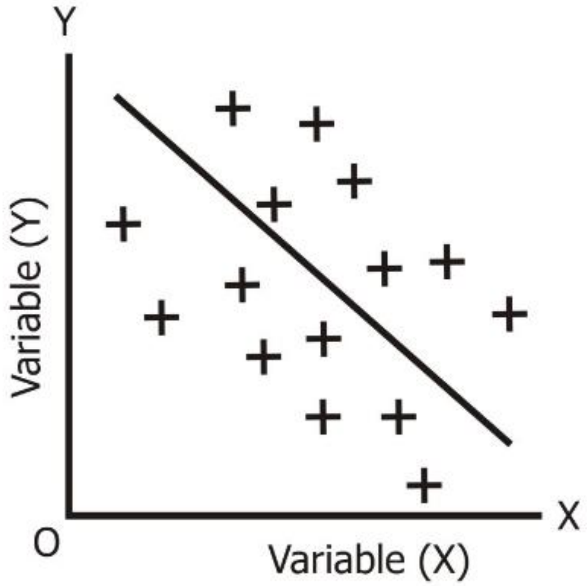
High Degree of Negative Correlation



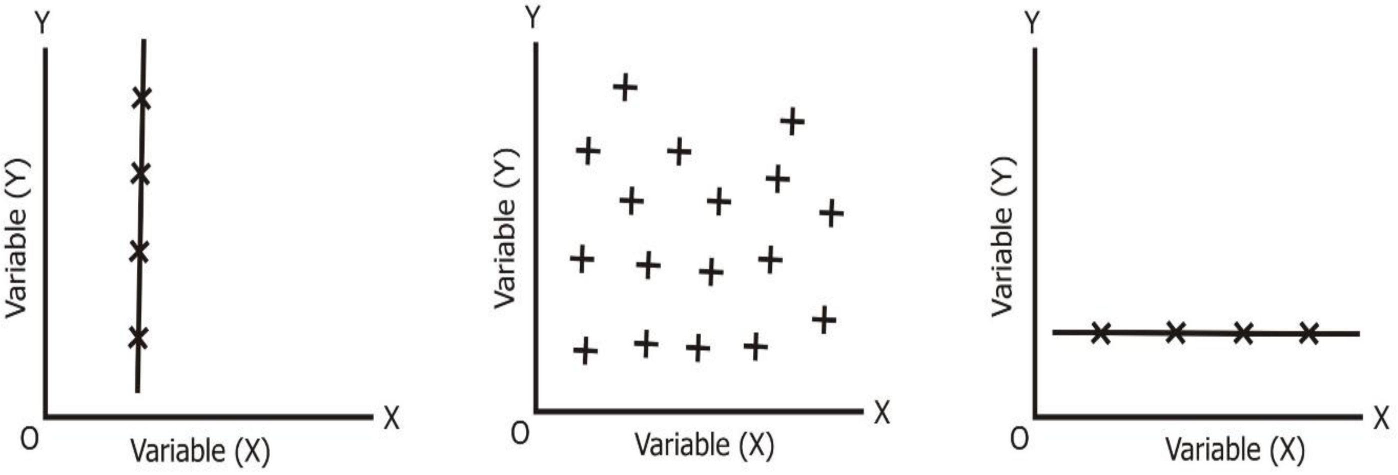
Partial Positive Correlation



Partial Negative Correlation



Lack of Correlation or No Correlation



Merits and Demerits of Scatter Diagram Method

Merits

1) This method is very simple to apply between two sets of variables.

2) Exclusive mathematical knowledge is not required.

3) Even though a few pairs of observation have more fluctuations, it will not influence much in the decision of the coefficient of correlation.

Demerits

1) Through this method one can get an idea about the relationship but not the exact value of coefficient of correlation.

2) Through this method it may happen that different people may give different opinions about the relationship among the variables.

2) Karl Pearson's Method

This method is the most accurate method to get degree among the variables numerically using moments of the variables. Also, this method is known as Karl Pearson's Product Moment Method. This method is mainly applicable for quantitative information. The following formulae can be used to establish coefficient of correlation among the data as and when required.

$$r = \frac{Covariance\ of\ (x, y)}{(S.D.\ of\ x).(S.D.\ of\ y)}$$

$$r=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{n.\sigma_x.\sigma_y} \text{ or } r=\frac{\sum xy-n\bar{x}\bar{y}}{n.\sigma_x.\sigma_y}$$

$$r=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2}\sqrt{\sum\left(y-\bar{y}\right)^2}}$$

$$r=\frac{n\sum xy-\sum x.\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

In case of large data, if change of origin and scale is used then above formula can be converted as follows:

Where,

$$u=\frac{x-A}{c_x} \text{ and } v=\frac{y-B}{c_y}$$

$$r=\frac{n\sum uv-\sum u.\sum v}{\sqrt{n\sum u^2-\left(\sum u\right)^2}\sqrt{n\sum v^2-\left(\sum v\right)^2}}$$

Merits and Demerits of Karl Pearson's Method

Merits

1) It gives precise value of the Coefficient of Correlation.

2) It gives direction as well as degree of relationship between two variables.

3) Based on this method the value of Coefficient of Correlation remains same even if different person has calculated the value.

Demerits

1) This method assumes that there is linear relationship between the variables regardless of the fact whether such relationship exists or not.

2) Compared to other methods this method requires more mathematical knowledge and is tough to apply.

3) This method is time consuming.

4) The Coefficient of Correlation is highly influenced by extreme pairs of observations.

3) Spearman's Rank Correlation Method

This method was suggested by Spearman who has used rank instead of the value of the data to establish correlation among the data. So, it is known as Spearman's Rank Correlation Method. This method is mainly applicable for qualitative information.

In this method both the series of data have been given rank separately after considering the highest data as first, next highest as second and so on. The following formula is used to find out coefficient of correlation among the data.

for non – repetitive ranking

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

Where,

d = Rank of X series ~ Rank of Y series

n = No. of pairs of observations

for repetitive ranking

$$r = 1 - \frac{6\left[\sum d^2 + \frac{m}{12}\left(m^2 - 1\right) + \frac{m}{12}\left(m^2 - 1\right) + \frac{m}{12}\left(m^2 - 1\right) + \ldots\right]}{n\left(n^2 - 1\right)}$$

Where,

d = Rank of X series ~ Rank of Y series

n = No. of pairs of observations

m = No. of time the item is repeat

Note: Here, $\frac{m}{12}\left(m^2 - 1\right)$ is correction factor which can be added as per the number of repetitions among the given data.

Merits and Demerits of Spearman's Rank Correlation Method

Merits

1) This method is easy to understand and easy to apply.
2) It is applicable between two associated attributes also.
3) It is applicable for qualitative data.
4) It will not be affected by extreme fluctuation because ranks are used to calculate the value of Coefficient of Correlation.

Demerits

1) This can be applied only on individual observation but not on grouped frequency distributions.
2) In ranking original values are replaced by their ranks so the results are approximate but not exact.
3) For large data this method becomes tedious.

- **Exercise**

✓ **Theoretical Questions**

1) What is correlation?
2) Explain different types of correlation.
3) Different methods to find coefficient of correlation.
4) Mention the use of correlation in practice.
5) Give uses of correlation in detail.
6) Give merits and demerits of scatter diagram.
7) Give merits and demerits of Karl Pearson's Method.
8) Give merits and demerits of Spearman's Rank Correlation Method.

MCQs

1) Which of the following is true for the coefficient of correlation?

   a) The coefficient of correlation is not dependent on the change of scale

   b) The coefficient of correlation is not dependent on the change of origin

   c) The coefficient of correlation is not dependent on both the change of scale and change of origin

   d) None of the above

   Answer: c

2) If the values of two variables move in the same direction, then

   a) The correlation is said to be non-linear

   b) The correlation is said to be linear

   c) The correlation is said to be negative

   d) The correlation is said to be positive

   Answer: d

3) If the values of two variables move in the opposite direction, then

   a) The correlation is said to be linear

   b) The correlation is said to be non-linear

   c) The correlation is said to be positive

   d) The correlation is said to be negative

   Answer: d

4) Which of the following statements is true about the correlational analysis between two sets of data?

   a) The correlational analysis between two sets of data is known as a simple correlation

   b) The correlational analysis between two sets of data is known as multiple correlation

   c) The correlational analysis between two sets of data is known as partial correlation

   d) None of the above

   Answer: a

5) Which one of the following statements about the correlation coefficient is correct?

   a) The correlation coefficient is unaffected by scale changes.

   b) Both the change of scale and the change of origin have no effect on the correlation coefficient.

   c) The correlation coefficient is unaffected by the change of origin.

   d) The correlation coefficient is affected by changes of origin and scale.

   Answer: c

6) Choose the correct option concerning the correlation analysis between 2 sets of data.

   a) Multiple correlation is a correlational analysis comparing two sets of data.

   b) A partial correlation is a correlational analysis comparing two sets of data.

   c) A simple correlation is a correlational analysis comparing two sets of data.

   d) None of the preceding.

   Answer: c

7) Which of the given plots is suitable for testing the linear relationship between a dependent and independent variable?

   a) Bar chart

   b) Scatter plot

   c) Histograms

   d) All of the above.

   Answer: b

8) The correlation for the values of two variables moving in the opposite direction is

   a) Positive

   b) Negative

   c) Linear

   d) No correlation.

   Answer: b

9) What is the type of coefficient of correlation if r = 0.81?

   a) Perfect Positive

   b) Perfect Negative

   c) Partial Positive

   d) Partial Negative

   Answer: c

10) Choose the correct example for positive correlation.

   a) Weight and income

   b) Price and demand

   c) The repayment period and EMI

   d) Income and expenditure

   Answer: d

BBA
SEMESTER-2
BUSINESS STATISTICS
BLOCK: 3

Authors' Name: Dr.Kalpesh Vankani, Assistant Professor, Christ College, Rajkot
Dr. Maulik Desai, Assistant Professor, K.S. School of Business Management,Ahmedabad

Review (Subject): Dr.Poonam Arora , Professor, SKIPS School of Business, Ahmedabad

Review (Language): Dr. Bhavna Trivedi , Assistant Professor, Dr. BAOU, Ahmedabad

Editor's Name: Prof. (Dr.) Manoj Shah
Professor and Director,
School of Commerce and Management,
Dr. Babasaheb Ambedkar Open University,
Ahmedabad.

Publisher's Name: Dr. Ajaysinh Jadeja
Registrar,
Dr. Babasaheb Ambedkar Open University
'JyotirmayParisar',Opp.ShriBalaji
Temple,Chharodi,Ahmedabad,382481,
Gujarat,India.

Edition: 2024 (First Edition)

ISBN: 978-93-5598-679-5

978-93-5598-679-5

UNIT-9 KARL PEARSON'S CO – EFFICIENT OF CORRELATION

9.1 Introduction

This method is the most accurate method to get degree among the variables numerically using moments of the variables. Also, this method is known as Karl Pearson's Product Moment Method. This method is mainly applicable for quantitative information.

9.2 Karl Pearson's Co – efficient of Correlation formula

The coefficient of correlation ('r') among the data can be found out as follows:

$$r = \frac{Covariance\ of\ (x, y)}{(S.D.\ of\ x).(S.D.\ of\ y)}$$

This can be simplified in different ways to solve the examples. Different forms of the formulae in detail have been mentioned in Unit – 8.

9.3 Skewness formula

To find out the direction and the extent of asymmetry in a series of statistical data skewness is used in the practice. The absolute measure of skewness tells us the extent of asymmetry and whether it is positive or negative.

Karl Pearson's Coefficient of Skewness can be found out as follows:

$$Sk = \frac{Mean - Mode}{S.D.} = \frac{\overline{x} - Mo}{\sigma}$$

Sometimes when for some distribution, it is difficult to determine mode precisely or it is ill defined, in this case skewness can be found out as follows.

$$Sk = \frac{3(Mean - Median)}{S.D.} = \frac{3\left(\overline{x} - Md\right)}{\sigma}$$

9.4 Application of Karl Pearson's Co – efficient of Correlation

- **Coefficient of Determination:**

The coefficient of determination (r^2) is the square of the coefficient of correlation. It is a more useful and readily comprehensible measure for indicating the percentage variation in the dependent variable which is accounted for by the independent variable. In other words, the coefficient of determination gives the ration of the explained variance to the total variance. Thus, coefficient of determination is explained as follows:

$$r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{\sum\left(y' - \bar{y}\right)^2}{\sum\left(y - \bar{y}\right)^2}$$

Where,

y = Original Series

$\bar{y}$ = Average of the series

y' = Estimated value of the y from regression line

- **Probable Error:**

Usually, we obtain coefficient of correlation of a sample drawn from a bivariate population. If different samples of the same size are drawn from a given population, we get different values of r. All these values of r differ from the actual value of the population correlation coefficient. The average of the absolute differences of correlation coefficients obtained from all possible samples and the population correlation coefficient is known as probable error of the correlation coefficient.

$$P.E. = \frac{0.6745\left(1 - r^2\right)}{\sqrt{n}}$$

The following rules can be applied to judge whether the correlation in the population is significant or not:

i. If $r <$ P.E. there is no evidence of correlation in the population i.e. the correlation in the population is not significant.

ii. If $r > 6$ (P.E.) there is evidence of significant correlation in the population.

iii. In other situation nothing can be stated with certainty.

iv. The probable error of correlation coefficient may be used to determine the limits within which the population correlation coefficient may be expected to lie r ± P.E. (r).

9.5 Correlation Co – efficient from bivariate frequency table

For bivariate table coefficient of correlation can be find out as follows:

Where,

$u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$

$$r = \frac{n\sum fuv - \sum ufu.\sum vfv}{\sqrt{n\sum u^2 fu - \left(\sum ufu\right)^2}\sqrt{n\sum v^2 fv - \left(\sum vfv\right)^2}}$$

9.6 Illustration

1) For a group of 10 items, $\sum x = 452$, $\sum x^2 = 24270$ and Mode = 43.7. Find Karl Pearson's Coefficient of Skewness.

Solution:

Here,

$$Mean = \bar{x} = \frac{\sum x}{n} = \frac{452}{10} = 45.2$$

$$S.D. = \sqrt{\left\{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2\right\}} = \sqrt{383.96} = 19.59$$

Now,

$$Sk = \frac{Mean - Mode}{S.D.} = \frac{\bar{x} - Mo}{\sigma} = \frac{45.2 - 43.7}{19.59} = 0.08$$

2) Calculate coefficient of correlation from the following data.

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Solution:

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$\left(X-\bar{X}\right)^2$	$\left(Y-\bar{Y}\right)^2$	$\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)$
64	66	-3	-2	9	4	6
65	67	-2	-1	4	1	2
66	65	-1	-3	1	9	3
67	68	0	0	0	0	0
68	70	1	2	1	4	2
69	68	2	0	4	0	0
70	72	3	4	9	16	12
469	**476**	**0**	**0**	**28**	**34**	**25**

Now $\bar{X}=\frac{\sum X}{n}=\frac{469}{7}=67$ and

$$\bar{Y}=\frac{\sum Y}{n}=\frac{476}{7}=68$$

Here both means are integers. So, following formula can be used

$$r=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2}\sqrt{\sum\left(y-\bar{y}\right)^2}}=\frac{25}{\sqrt{28}\sqrt{34}}=0.81$$

3) Calculate coefficient of correlation from the following data.

X	39	65	62	90	82	75	25	98	36	78
Y	47	53	58	86	62	68	60	91	51	84

Solution:

X	Y	X^2	Y^2	$X.Y$
39	47	1521	2209	1833
65	53	4225	2809	3445
62	58	3844	3364	3596
90	86	8100	7396	7740

82	62	6724	3844	5084
75	68	5625	4624	5100
25	60	625	3600	1500
98	91	9604	8281	8918
36	51	1296	2601	1836
78	84	6084	7056	6552
650	**660**	**47648**	**45784**	**45604**

Here coefficient of correlation can be found out using following formula

$$r = \frac{n\sum xy - \sum x.\sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

$$\therefore r = \frac{10(45604) - (650).(660)}{\sqrt{10(47648) - (650)^2}\sqrt{10(45784) - (660)^2}} = 0.7804$$

4) Calculate suitable coefficient of correlation for the following data.

Price	15	18	20	24	30	35	40	50
Supply	85	93	95	105	120	130	150	160

Solution:

For large data one can use step deviation method to calculate coefficient of correlation as follows:

Where,

$$u = \frac{x - A}{c_x} = \frac{x - 29}{1} \text{ and } v = \frac{y - B}{c_y} = \frac{y - 119}{1}$$

$$r = \frac{n\sum uv - \sum u.\sum v}{\sqrt{n\sum u^2 - \left(\sum u\right)^2}\sqrt{n\sum v^2 - \left(\sum v\right)^2}}$$

$$\therefore r = \frac{8(2317)-(0)(-14)}{\sqrt{8(1022)-(0)^2}\sqrt{8(5368)-(-14)^2}} = 0.9917$$

5) The following table shows the scores in a class test of 67 students of different age groups. Find the coefficient of correlation between age and scores of the test.

Scores	**Age in Years**				**Total**
	18	**19**	**20**	**21**	
200 – 250	4	4	2	1	**11**
250 – 300	3	5	4	2	**14**
300 – 350	2	6	8	5	**21**
350 – 400	1	4	6	10	**21**
Total	**10**	**19**	**20**	**18**	**n = 67**

Solution:

For bivariate data following calculation is required

y	**x**									
	18	**19**	**20**	**21**	**fy = fv**	**M.V. of y**	**v**	**vfv**	**v^2fv**	**fuv**
200 – 250	(16) 4	(8) 4	(0) 2	(-2) 1	**11**	225	-2	-22	44	22
250 – 300	(6) 3	(5) 5	(0) 4	(-2) 2	**14**	275	-1	-14	14	9
300 – 350	(0) 2	(0) 6	(0) 8	(0) 5	**21**	325	0	0	0	0
350 – 400	(-2) 1	(-4) 4	(0) 6	(10) 10	**21**	375	1	21	21	4
fx = fu	**10**	**19**	**20**	**18**	**n = 67**			$\sum$vfv = -15	$\sum v^2$fv = 79	$\sum$fuv = 35
M.V. of x	18	19	20	21						
u	-2	-1	0	1						
ufu	-20	-19	0	18	$\sum$ ufu = - 21					
u^2fu	40	19	0	18	$\sum u^2$fu = 77					
fuv	20	9	0	6	$\sum$ fuv = 35					

Where,

$$u = \frac{x - A}{c_x} = \frac{x - 20}{1} \text{ and } v = \frac{y - B}{c_y} = \frac{y - 325}{50}$$

$$r = \frac{n\sum fuv - \sum ufu.\sum vfv}{\sqrt{n\sum u^2 fu - \left(\sum ufu\right)^2}\sqrt{n\sum v^2 fv - \left(\sum vfv\right)^2}}$$

$$\therefore r = \frac{67(35) - (-21)(-15)}{\sqrt{67(77) - (-21)^2}\sqrt{67(79) - (-15)^2}} = 0.42$$

6) The following data are obtained for two variables x and y:

n = 30, ∑ = 120, ∑ xy = 356, ∑ x^2 = 600, ∑ y = 90, ∑ y^2 = 250

However later on it was observed that two pairs were wrongly taken as (8, 10) and (12, 7) instead of (8, 12) and (10, 8). Find the correct value of the correlation coefficient.

Solution:

Correct Values: (8, 12) and (10, 8)

Wrong Values: (8, 10) and (12, 7)

Corrected totals are as follows:

n = 30 – 1 – 1 + 1 + 1 = 30

∑ x = 120 – 8 – 12 + 8 + 10 = 118

∑ y = 90 -10 – 7 + 12 + 8 = 93

∑ x^2 = 600 – 8^2 – 12^2 + 8^2 + 10^2 = 556

∑ y^2 = 250 – 102 – 72 + 122 + 82 = 309

∑ xy = 356 – (8x10) – (12x7) + (8x12) + (10x8) = 368

Now corrected coefficient of correlation is

$$r = \frac{n\sum xy - \sum x.\sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

$$\therefore r=\frac{30(368)-(118).(93)}{\sqrt{30(556)-(118)^2}\sqrt{30(309)-(93)^2}}=0.05$$

7) The correlation coefficient obtained from a sample of 16 pairs of observations drawn from a population is 0.7. Calculate the probable error of the correlation coefficient and interpret it. Also find the limits of the population correlation coefficient.

Solution:

Here,

$$P.E.=\frac{0.6745\left(1-r^2\right)}{\sqrt{n}}=\frac{036745\left(1-0.7^2\right)}{\sqrt{16}}=0.086$$

Also, r = 0.7 > 6(P.E.) = 6(0.086) = 0.516

So, there is a significant correlation between two variables.

The probable limits of the population correlation coefficient is

r ± P.E. (r). = 0.7 ± 0.086 = 0.614 to 0.786

8) The correlation coefficient between two variables x and y is 0.48 and the covariance between them is 36. If the variance of x is 16, find S.D. of y.

Solution:

$$r=\frac{Covariance\ of\ (x,y)}{(S.D.\ of\ x).(S.D.\ of\ y)}$$

$$\therefore 0.48=\frac{36}{(4).(S.D.\ of\ y)}\Rightarrow S.D.\ of\ y=18.75$$

9) The correlation coefficient between two variables is 0.07, calculate value of the coefficient of determination. Also interpret the same.

Solution:

coefficient of determination (r^2) = (coefficient of correlation)^2 = 0.49

So, 49 % of the variation is explained by the independent variables among the dependent variables.

❖ Exercise

✓ Theoretical Questions

1) Explain Karl Pearson's Coefficient of Correlation Method in detail.
2) Explain coefficient of determination with proper examples.
3) Explain probable error.
4) How to find out skewness through Karl Pearson's method?

✓ MCQs

1) If the correlation between X and Y is 0.3, then correlation between 2X and 3Y is

   a) 0.3
   b) 0.4
   c) 0.2
   d) None of the above

   Answer: a

2) If the correlation coefficient between X and Y is 0.8 and covariance is 121 and the variance of Y is 64, then variance of X will be

   a) 357.59
   b) 1237
   c) 158
   d) 18.91

   Answer: a

3) If the covariance between X and Y is 12, variance of X is 64 and variance of Y is 36, then what is the value of coefficient of correlation?

   a) 1 / 4
   b) 1 / 3
   c) 1 / 2
   d) 2 / 3

   Answer: a

4) For two correlated variables X and Y, if coefficient of correlation between X and Y is 0.8014, variance of X and Y are 16 and 25 respectively. Then the covariance between X and Y is

   a) 162.08
   b) 16.028
   c) 160.28

d) 16.208

Answer: b

5) The Karl Pearson's Coefficient of Correlation Method between two variables X and Y can be find out by

a) $r = \dfrac{Covariance\ of\ (x, y)}{(S.D.\ of\ x).(S.D.\ of\ y)}$

b) $r = \dfrac{(S.D.\ of\ x).(S.D.\ of\ y)}{Covariance\ of\ (x, y)}$

c) $r = Covariance\ of\ (x, y) * (S.D.\ of\ x).(S.D.\ of\ y)$

d) $r = \dfrac{Covariance\ of\ (x, y)}{(\text{Variance}\ of\ x).(\text{variance}\ of\ y)}$

Answer: a

6) The coefficient of correlation is independent of

a) Change of scale only

b) Change of origin only

c) Neither change of origin nor change of scale

d) Both change of origin and change of scale

Answer: d

7) If the coefficient of correlation r = 0, then both the variables are

a) No linear relation between them

b) Variables are uncorrelated

c) Independent

d) All of the above

Answer: d

8) The range of the coefficient of correlation is

a) [0, ∞)

b) R

c) [- 1, 1]

d) None of the above

Answer: c

9) If coefficient of correlation is 0.5 then what is the value of the coefficient of determination

a) 0.25

b) 0.05

c) 0.5

d) – 0.5

Answer: a

10) For two variables, r = 0.8 and the probable error of r is 0.08. Find the number of pairs of observations.

a) 6

b) 7

c) 8

d) 9

Answer: d

✓ **Practical Examples**

1) The mean of a certain distribution is 50, its standard deviation is 15 and coefficient of skewness is – 1. Find the median.

Answer: Median = 65

2) Calculate Karl Pearson's coefficient of skewness

x:	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5
f:	28	42	54	108	129	61	45	35

Answer: Mean = 30.52, Mode = 31.18, S.D. = 9.003, Sk = - 0.072

3) The data relating to price and quantity in respect of a given commodity are as under:

Price	2	3	6	5	4	3	5	7	8	7
Quantity	6	5	4	5	7	10	9	7	8	9

Calculate Karl Pearson's coefficient of correlation between price and quantity.

Answer: r = 0.14

4) Calculate correlation coefficient between X and Y for the following data:

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

Answer: 0.95

5) Calculate Karl Pearson’s coefficient of correlation from the data of price and demand

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

Answer: r = - 0.954

6) Calculate coefficient of correlation between employment and sales as given below:

Employment (‘000 person)	22	31	90	82	43	62	59	16	61	46	35	50
Sales (‘000 rupees)	250	980	980	850	710	280	530	180	670	420	190	460

Answer: 0.58

7) Calculate the Pearson’s coefficient of correlation from the following data.

X	43	44	46	40	44	42	45	42	38	40	42	57
Y	29	31	19	18	19	27	27	29	41	30	26	10

Answer: r = - 0.73

8) Compute Karl Pearson’s coefficient of correlation in the following series relating to price and supply of commodity.

Price (₹)	70	75	80	85	90	95	100	105	110	115
Supply (Qua.)	30	29	25	25	22	20	18	18	16	10

Answer: r = - 0.97918

9) Calculate Karl Pearson’s coefficient of correlation of the following series:

Price (in ₹)	**Demand (in kg.)**
110 – 111	600
111 – 112	610
112 – 113	640
113 – 114	680

114 – 115	700
115 – 116	780
116 – 117	830
117 – 118	900
118 – 119	1000

Also calculate the probable error of the correlation coefficient.

Answer: r = 0.9708, P.E. = 0.02241

10) The following table gives the distribution of the age in years of husband and wives of 100 couples:

Age of Wives	Age of Husband			
	20 - 25	**25 - 30**	**30 - 35**	**35 - 40**
15 – 20	20	10	3	2
20 – 25	4	28	6	4
25 – 30	-	5	11	-
30 – 35	-	-	2	-
35 – 40	-	-	-	5

Calculate the coefficient of correlation between the age of husband and age of the wife.

Answer: r = 0.61

11) Calculate coefficient of correlation and its probable error for the following data:

Y	X				
	0 – 500	**500 – 1000**	**1000 – 1500**	**1500 – 2000**	**2000 – 2500**
0 – 200	12	6	-	-	-
200 – 400	2	18	4	2	1
400 – 600	-	4	7	3	-
600 – 800	-	1	-	2	1
800 – 1000	-	-	1	2	3

Answer: r = 0.76, P.E. = 0.1645

12) Calculate Karl Pearson's coefficient of correlation from the data given below:

Marks	Age in years				
	18	**19**	**20**	**21**	**22**

20 – 25	3	2	-	-	-
15 – 20	-	5	4	-	-
10 – 15	-	-	7	10	-
5 – 10	-	-	-	3	2
0 – 5	-	-	-	3	1

What is the value of the coefficient of determination.

Answer: r = - 0.84, coefficient of determination = 0.7056

UNIT-10 SPEARMAN'S RANK CORRELATION

10.1 Introduction

This method was suggested by Spearman who has used rank instead of the value of the data to establish correlation among the data. So, it is known as Spearman's Rank Correlation Method. This method is mainly applicable for qualitative information.

10.2 Concept of Rank Correlation

In this method instead of actual values rank has been used to get coefficient of correlation. Also, if the data consists of qualitative information, then also this method is applicable to get coefficient of correlation.

10.3 Rank Correlation formula

This method was suggested by scientist Spearman which is used for the data where ranks have been used.

In this method both the series of data have been given rank separately after considering the highest data as first, next highest as second and so on. The following formula is used to find out coefficient of correlation among the data.

for no–repetitive ranking

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

Where,
d = Rank of X series ~ Rank of Y series
n = No. of pairs of observations

for repetitive ranking

$$r = 1 - \frac{6\left[\sum d^2 + \frac{m}{12}\left(m^2 - 1\right) + \frac{m}{12}\left(m^2 - 1\right) + \frac{m}{12}\left(m^2 - 1\right) + \ldots\right]}{n\left(n^2 - 1\right)}$$

Where,
d = Rank of X series ~ Rank of Y series

n = No. of pairs of observations
m = No. of time the item is repeat

Note: Here, $\frac{m}{12}\left(m^2 - 1\right)$ is correction factor which can be added as per the number of repetitions among the given data.

10.4 Advantages and Disadvantages of Spearman's Rank Correlation

We have already discussed advantages and disadvantages of Spearman's Rank Correlation in Unit – 8. So, one can refer to the same.

10.5 Illustration

1) The ranks of 15 students in Statistics and Mathematics are given below. The data is as follows:

Statistics	1	7	2	9	12	8	6	3	13	15	14	10	11	4	5
Mathematics	2	9	1	7	15	8	5	3	13	14	11	10	12	6	4

Find rank correlation coefficient.

Solution:

Rank in Statistics (R_x)	**Rank in Mathematics (R_y)**	**$d = R_x - R_y$**	**d^2**
1	2	-1	1
7	9	-2	4
2	1	1	1
9	7	2	4
12	15	-3	9
8	8	0	0
6	5	1	1
3	3	0	0
13	13	0	0
15	14	1	1
14	11	3	9
10	10	0	0
11	12	-1	1
4	6	-2	4
5	4	1	1
Total			36

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(36)}{15\left(15^2 - 1\right)} = 0.9357$$

2) Find coefficient of rank correlation.

X	28	27	26	35	39	42	39	37	32	22
Y	40	42	38	49	40	50	38	44	45	36

Solution:

First of all, give separate ranking for both the series considering highest value as first rank and so on.

X	**Y**	**R_x**	**R_y**	**d = R_x - R_y**	**d^2**
28	40	7	6.5*	0.5	0.25
27	42	8	5	3	9
26	38	9	8.5*	0.5	0.25
35	49	5	2	3	9
39	40	2.5*	6.5*	-4	16
42	50	1	1	0	0
39	38	2.5*	8.5*	-6	36
37	44	4	4	0	0
32	45	6	3	3	9
22	36	10	10	0	0
Total				0	$\sum d^2 = 79.5$

** Here 39, 40 and 38 repeats twice so in continuation we have given rank to them by taking average of the actual rank for 39 as $\left(\frac{2+3}{2} = 2.5\right)$, for 40 as $\left(\frac{6+7}{2} = 6.5\right)$ and for 38 as $\left(\frac{8+9}{2} = 8.5\right)$

So, there are total three repetitions in the ranking among them, and they repeat twice. So, in the formula three correction factors should be added where for one value of m = 2 for each of the correction factor.

For repetitive ranking

$$r=1-\frac{6\left[\sum d^2+\frac{m}{12}\left(m^2-1\right)+\frac{m}{12}\left(m^2-1\right)+\frac{m}{12}\left(m^2-1\right)\right]}{n\left(n^2-1\right)}$$

$$\therefore r=1-\frac{6\left[79.5+\frac{2}{12}\left(2^2-1\right)+\frac{2}{12}\left(2^2-1\right)+\frac{2}{12}\left(2^2-1\right)\right]}{10\left(10^2-1\right)}=0.5091$$

3) Calculate Spearman's Rank Correlation from the following data.

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

Solution:

We can give separate ranking to both series as following, after considering highest as first.

X	**Y**	**R_x**	**R_y**	**d = R_x - R_y**	**d^2**
48	13	3	5.5**	- 2.5	6.25
33	13	5	5.5**	- 0.5	0.25
40	24	4	1	3	9
9	6	10	8.5**	1.5	2.25
16	15	8*	4	4	16
16	4	8*	10	- 2	4
65	20	1	2	- 1	1
24	9	6	7	- 1	1
16	6	8*	8.5**	- 0.5	0.25
57	19	2	3	- 1	1
Total					$\sum d^2 = 41$

* Here 16 is repeated thrice so in continuation we have given rank to them by taking average of the actual rank as $\left(\frac{7+8+9}{3}=8\right)$

** Here 13 and 6 are repeated twice so in continuation we have given rank to them by taking average of the actual rank for 13 as $\left(\frac{5+6}{2}=5.5\right)$ and for 6 as $\left(\frac{8+9}{2}=8.5\right)$

So, there are total three repetitions in the ranking among them. One observation repeats thrice, and the other two observations repeat twice. So, in the formula three correction factors should be added where for one value of m = 3 and for the remaining two value of the m = 2. Which is shown as follows:

For repetitive ranking

$$r = 1 - \frac{6\left[\sum d^2 + \frac{m}{12}\left(m^2-1\right) + \frac{m}{12}\left(m^2-1\right) + \frac{m}{12}\left(m^2-1\right)\right]}{n\left(n^2-1\right)}$$

$$\therefore r = 1 - \frac{6\left[41 + \frac{3}{12}\left(3^2-1\right) + \frac{2}{12}\left(2^2-1\right) + \frac{2}{12}\left(2^2-1\right)\right]}{10\left(10^2-1\right)} = 0.7333$$

4) Ten competitors in a beauty contest are ranked by three judges in the following order:

Judge 1	1	5	4	8	9	6	10	7	3	2
Judge 2	4	8	7	6	5	9	10	3	2	1
Judge 3	6	7	8	1	5	10	9	2	3	4

Use the rank correlation coefficient to discuss which pair of judges has the nearest approach in beauty.

Solution:

We can the given data as follows

R_1	R_2	R_3	$d_{12} = R_1 - R_2$	$d_{23} = R_2 - R_3$	$d_{31} = R_3 - R_1$	d^2_{12}	d^2_{23}	d^2_{31}
1	4	6	-3	-2	5	9	4	25
5	8	7	-3	1	2	9	1	4
4	7	8	-3	-1	4	9	1	16

8	6	1	2	5	-7	4	25	49
9	5	5	4	0	-4	16	0	16
6	9	10	-3	-1	4	9	1	16
10	10	9	0	1	-1	0	1	1
7	3	2	4	1	-5	16	1	25
3	2	3	1	-1	0	1	1	0
2	1	4	1	-3	2	1	9	4
Total						**$\sum d^2_{12}$ = 74**	**$\sum d^2_{23}$ = 44**	**$\sum d^2_{31}$** = 156

Now,

$$r_{12} = 1 - \frac{6\sum d_{12}^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(74)}{10\left(10^2 - 1\right)} = 0.5515$$

$$r_{23} = 1 - \frac{6\sum d_{23}^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(44)}{10\left(10^2 - 1\right)} = 0.7333$$

$$r_{31} = 1 - \frac{6\sum d_{31}^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(156)}{10\left(10^2 - 1\right)} = 0.0545$$

From the above rank correlation between the different pairs of judges we can conclude that judge 2 and judge 3 have the nearest approach in beauty.

5) The coefficient of rank correlation of the marks obtained by 10 students in two particular subjects was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. What should be the correct value of coefficient of rank correlation?

Solution:

Here, $r = 0.5, n = 10$

Now, $r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$

$$\therefore 0.5 = 1 - \frac{6\sum d^2}{10\left(10^2 - 1\right)} \Rightarrow \sum d^2 = 82.5$$

Here one of the differences in ranks in two subjects was wrongly taken as 3 instead of 7. So, the corrected total is

$$\therefore \sum d^2 = 82.5 - 3^2 + 7^2 = 122.5$$

So, corrected rank correlation is

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6\left(122.5\right)}{10\left(10^2 - 1\right)} = 0.2576$$

6) In a bivariate sample the sum of squares of differences between the ranks of observed values of two variables is 231 and the correlation coefficient between them is – 0.4 then find the number of pairs.

Solution:

Here, $r = -0.4, \sum d^2 = 231, n = \left(?\right)$

Now, $r = 1 - \dfrac{6\sum d^2}{n\left(n^2 - 1\right)}$

$$\therefore -0.4 = 1 - \frac{6\left(231\right)}{n\left(n^2 - 1\right)}$$

$$\therefore n\left(n^2 - 1\right) = 990$$

$$\therefore \left(n - 1\right)n\left(n + 1\right) = 9X10X11$$

By comparing the corresponding term n = 10.

- **Exercise**

✓ **Theoretical Questions**

1) Explain Spearman's Rank Correlation method in detail.
2) Give advantages of Rank Correlation method.
3) Give disadvantages of Rank Correlation method.
4) Give the difference between Pearson's Correlation method and Spearman's Rank Correlation method.

✓ **MCQs**

1) The rank correlation coefficient between marks obtained by 10 students in English and Mathematics was found to be 0.5. Find the sum of squares of ranks.
   a) 81.5
   b) 82.5
   c) 83.5
   d) 84.5

   Answer: b

2) For Spearman's rank correlation, if the correlation coefficient is 0.7 and $\sum d_i^2$ = 49.5 then the value of sample size n is
   a) 99
   b) 20
   c) 10
   d) 990

   Answer: c

3) For two variables X and Y, the following observations are tabulated. The spearman's correlation coefficient is

X	4	3	1
Y	6	7	10

   a) 0
   b) 1
   c) – 1
   d) ∞

   Answer: c

4) If the difference between the rank of the 4 observations is 2.5, 0.5, -1.5, -1.5 then Spearman's rank coefficient of correlation is
   a) – 0.2
   b) 0.1
   c) – 0.1
   d) 0.2

   Answer: - 0.1

5) Spearman's rank correlation coefficient is given by

   a) $r = 1 + \frac{6\sum d^2}{n\left(n^2 - 1\right)}$

b) $r = 1 - \frac{6\sum d^2}{n\left(n^2+1\right)}$

c) $r = 1 - \frac{6\sum d}{n\left(n^2-1\right)}$

d) $r = 1 - \frac{6\sum d^2}{n\left(n^2-1\right)}$

Answer: d

6) If in a series, there are m items whose ranks are common then which factor is added for correlation for each repeating value in both series

   a) $\frac{m}{12}$

   b) $\frac{m}{12}\left(m^2-1\right)$

   c) $\frac{1}{12}\left(m^2-1\right)$

   d) $\frac{m}{12}\left(m^2+1\right)$

   Answer: b

7) If the sum of the squares of difference of ranks of 6 candidates in two criteria is 21, the rank correlation coefficient is

   a) 0.5

   b) 0.6

   c) 0.4

   d) 0.7

   Answer: c

8) The rank correlation method was offered by

   a) Likert

   b) Wilcoxon

   c) Pearson

   d) Spearman

   Answer: d

9) The rank correlation coefficient for 10 pairs is 0.3. Later, it was found that the difference in ranks of one pair was misread as 9 instead of 6. Find the correct value of correlation coefficient.

   a) 1.57

   b) 0.57

   c) – 1.57

   d) – 0.57

   Answer: b

10) If the sum of squares of differences in ranks of two variables x and y is 126 and the correlation coefficient is – 0.5, find the number of pairs.

   a) 5

   b) 6

   c) 7

   d) 8

   Answer: d

✓ **Practical Examples**

1) The ranking of 10 trainees at the beginning and at the end of a certain course are given below:

Trainees **Rank**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Before	1	6	3	9	5	2	7	10	8	4
After	6	8	3	7	2	1	5	9	4	10

Obtain the Spearman's coefficient of correlation.

Answer: 0.394

2) Find rank correlation coefficient from the following data.

X	70	40	42	48	35	38	40	45	70	55	51
Y	120	125	124	120	120	130	128	122	110	116	118

Answer: - 0.73

3) From the following data calculate the coefficient of rank correlation between X and Y.

X	43	32	55	49	60	43	37	43	49	10	20
Y	60	40	30	70	20	30	50	72	60	45	25

Answer: - 0.11

4) The coefficient of rank correlation of the marks obtained by 10 students in two subjects was found to be 0.5. It was then detected that the difference in ranks in the two subjects obtained by one of the students

was wrongly taken as 3 in place of 7. What should be the correct rank correlation?

Answer: 0.26

5) The sum of the squares of the difference in the difference in the ranks of n pairs of observation is known to be 126 whereas the coefficient of rank correlation equals to – 0.5. what will be the value of n?

Answer: n = 8

6) Value of the Spearman's rank correlation coefficient for a certain number of pairs of observation was found to be $\frac{2}{3}$. The sum of the squares of difference between corresponding ranks was 55. Find the number of pairs.

Answer: 10

7) Ten competitors in an intelligence test are ranked by three judges in the following order:

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

Use the method of Rank Correlation to determine which pair of judges has the nearest approach to common liking in intelligence.

Answer:

$$r_{12} = -0.21, r_{23} = -0.3, r_{13} = 0.64$$

1st and 3rd judge have most common liking among the competitors.

UNIT-11 REGRESSION ANALYSIS

11.1 Introduction

Regression analysis is the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables. The word Regression was used by Sir Francis Galton in the nineteenth century. The most elementary regression model is called simple regression or bivariate regression involving two variables in which one variable is predicted by another variable. In simple regression, the variable to be predicted is called the dependent variable and is designated as y. The predictor is called the independent variable, or explanatory variable, and is designated as x. In simple regression analysis, only a straight-line relationship between two variables is examined. Nonlinear relationships and regression models with more than one independent variable can be explored by using multiple regression models.

Regression can be defined as follows:

"Regression is the functional relationship between two variables which have cause and effect relationship, where with the help of independent variable the value of dependent variable can be established."

11.2 Utility of Regression Analysis

1) It helps us in finding the value of the dependent variable with the help of independent variable.
2) One can also find out the error of the regression line.
3) It is highly used in economics and business research.
4) It is useful in business forecasting as a tool.
5) It is widely used in mathematical statistics in estimation of demand, supply, etc.

11.3 Equation of straight line

- ➢ **Equation of Regression Line Y on X**

  $Y = a + b_{yx} . X$

  Where,

  X = Independent Variable

  Y = Dependent Variable

a = Constant Term

b_{yx} = Regression coefficient Y on X

Also,

$$a = \bar{y} - b_{yx}.\bar{x}$$

Where $\bar{x} = \frac{\sum x}{n}$ and $\bar{y} = \frac{\sum y}{n}$

Also,

$$b_{yx} = \frac{Cov(x,y)}{\text{Variance of X}} = r\frac{s_y}{s_x}$$

$$\therefore b_{yx} = \frac{\sum\left(x-\bar{x}\right).\left(y-\bar{y}\right)}{\sum\left(x-\bar{x}\right)^2}$$

$$\therefore b_{yx} = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

If u = x – A and v = y – B (where A and B are assumed mean)

$$\therefore b_{yx} = \frac{n\sum uv - \sum u\sum v}{n\sum u^2 - (\Sigma u)^2}$$

If $u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$

$$\therefore b_{yx} = \frac{n\sum uv - \sum u\sum v}{n\sum u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

For bivariate data

$$\therefore b_{yx} = \frac{n\sum fuv - \sum ufu\sum vfv}{n\sum u^2 fu - (\Sigma ufu)^2} \times \frac{c_y}{c_x}$$

$$a = \bar{y} - b_{yx}.\bar{x}$$

Where $\bar{x} = A + \frac{\sum ufu}{n} \times c_x$ and $\bar{y} = B + \frac{\sum vfv}{n} \times c_y$

- **Equation of Regression Line X on Y**

$X = a' + b_{xy} . Y$

Where,

X = Dependent Variable

Y = Independent Variable

A = Constant Term

b_{xy} = Regression coefficient X on Y

Also,

$$a' = \bar{x} - b_{xy} . \bar{y}$$

Where $\bar{x} = \frac{\sum x}{n}$ and $\bar{y} = \frac{\sum y}{n}$

Also,

$$b_{xy} = \frac{Cov(x,y)}{\text{Variance of Y}} = r\frac{s_x}{s_y}$$

$$\therefore b_{xy} = \frac{\sum\left(x-\bar{x}\right).\left(y-\bar{y}\right)}{\sum\left(y-\bar{y}\right)^2}$$

$$\therefore b_{xy} = \frac{n\sum xy - \sum x\sum y}{n\sum y^2 - \left(\sum y\right)^2}$$

If u = x – A and v = y – B (where A and B are assumed mean)

$$\therefore b_{xy} = \frac{n\sum uv - \sum u\sum v}{n\sum v^2 - \left(\Sigma v\right)^2}$$

If $u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$

$$\therefore b_{xy} = \frac{n\sum uv - \sum u\sum v}{n\sum v^2 - \left(\Sigma v\right)^2} \times \frac{c_x}{c_y}$$

For bivariate data

$$\therefore b_{xy} = \frac{n\sum fuv - \sum ufu \sum vfv}{n\sum v^2 fv - \left(\sum vfv\right)^2} \times \frac{c_x}{c_y}$$

$$a' = \bar{x} - b_{xy}.\bar{y}$$

Where $\bar{x} = A + \frac{\sum ufu}{n} \times c_x$ and $\bar{y} = B + \frac{\sum vfv}{n} \times c_y$

- **Properties of Regression Coefficients**
  1) The product of regression coefficient is the square of the correlation coefficient.
  2) b_{yx}, b_{xy} and r have always had the same sign.
  3) If two variables have a perfect relationship, then one regression coefficient is reciprocal of the other.
  4) The product of two regression coefficient is r2 which cannot exceed 1.
  5) The regression coefficients are independent of change of origin but not of scale.

11.4 Difference Between Correlation and Regression

No.	Description	Correlation	Regression
1	Purpose	Measures the strength and direction of a linear relationship between two variables.	Focuses on predicting the value of one variable based on the value of another.
2	Concept	It indicates whether and how strongly pairs of variables are related. It produces a single value (called the correlation coefficient), which ranges from -1 to +1.	It involves finding an equation that best describes the relationship between the variables, often in the form of a straight line (in simple linear regression)
3	Symmetry vs. Asymmetry	It is symmetric, meaning the correlation between X and Y is the same as between Y and X.	It is asymmetric. The regression of Y on X is not necessarily the same as the regression of X on Y because regression attempts to model how one variable predicts or depends on the other.
4	Variables	Examines the relationship between two variables without	Involves a distinction between independent (predictor) and dependent

		making any distinction between dependent and independent variables. Both variables are treated equally.	(outcome) variables. One variable is being predicted (dependent), while the other is used to make predictions (independent).
5	Interpretation	It is concerned only with the strength and direction of the relationship, not causality. A strong correlation doesn't imply that one variable causes the other.	It is used to understand the causal relationship (or at least the predictive relationship) between the variables. For example, it tells you how much Y changes when X changes.
6	Result	The output is a correlation coefficient, which quantifies the degree of the relationship between the variables.	The output is an equation (e.g., a regression line in simple linear regression) that can be used for predicting future values.
7	Use	Used when you want to measure the strength and direction of a relationship without making predictions (e.g., checking if height and weight are related).	Used when you want to predict the value of one variable based on the value of another (e.g., predicting a person's weight based on their height).
8	Summary	Correlation tells you if and how two variables are related.	Regression tells you how one variable can be used to predict another.

11.5 Illustration

1) Find regression coefficients from the following. From the value of regression coefficient find the value of the coefficient of correlation.

X	21	22	23	24	25	26	27	28	29	30
Y	17	19	19	20	23	24	27	26	28	27

Solution:

X	**Y**	**u = X - 25**	**v = y – 23**	**u2**	**v2**	**uv**
21	17	-4	-6	16	36	24
22	19	-3	-4	9	16	12
23	19	-2	-4	4	16	8
24	20	-1	-3	1	9	3

25	23	0	0	0	0	0
26	24	1	1	1	1	1
27	27	2	4	4	16	8
28	26	3	3	9	9	9
29	28	4	5	16	25	20
30	27	5	4	25	16	20
Total		5	0	85	144	105

Here u = x – A = x – 25 and v = y – B = y – 23 (where A and B are assumed mean)

For regression coefficient Y on X

$$\therefore b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - (\Sigma u)^2} = \frac{10(105) - 5(0)}{10(85) - (5)^2} = 1.27$$

For regression coefficient X on Y

$$\therefore b_{xy} = \frac{n\sum uv - \sum u \sum v}{n\sum v^2 - (\Sigma v)^2} = \frac{10(105) - 5(0)}{10(144) - (0)^2} = 0.73$$

Now,

$$r = \sqrt{b_{yx} b_{xy}} = \sqrt{(1.27)(0.73)} = 0.9629$$

2) Obtain both the regression lines from the following data:

X	1	5	3	2	1	2	7	3
Y	6	1	0	0	1	2	1	5

Also, calculate value of X when Y = 3 and Y when X = 4.

Solution:

Here n = 8

X	**Y**	**X^2**	**Y^2**	**XY**
1	6	1	36	6
5	1	25	1	5
3	0	9	0	0
2	0	4	0	0

1	1	1	1	1
2	2	4	4	4
7	1	49	1	7
3	5	9	25	15
$\sum X = 24$	**$\sum Y = 16$**	**$\sum X^2 = 102$**	**$\sum Y^2 = 68$**	**$\sum XY = 38$**

- Equation of Regression Line Y on X

$$Y = a + b_{yx} . X$$

Where,

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{8(38)-(24)(16)}{8(102)-(24)^2} = -0.33$$

Also,

$$a = \bar{y} - b_{yx}.\bar{x}$$

Where $\bar{x} = \frac{\sum x}{n} = \frac{24}{8} = 3$ and $\bar{y} = \frac{\sum y}{n} = \frac{16}{8} = 2$

$$\therefore a = \bar{y} - b_{yx}.\bar{x} = 2 - (-0.33)(3) = 2.99$$

So, regression equation Y on X is Y = 2.99 – 0.33.X -------- (1)

From equation (1) when X = 4 then Y = 2.99 – 0.33.(4) = 1.67

- **Equation of Regression Line X on Y**

$$X = a' + b_{xy} . Y$$

Where,

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2} = \frac{8(38)-(24)(16)}{8(68)-(16)^2} = -0.28$$

Also,

$$a' = \bar{x} - b_{xy}.\bar{y}$$

Where $\bar{x} = \frac{\sum x}{n} = \frac{24}{8} = 3$ and $\bar{y} = \frac{\sum y}{n} = \frac{16}{8} = 2$

$$\therefore a' = \bar{x} - b_{xy} . \bar{y} = 3 - (-0.28)(2) = 3.56$$

So, regression equation X on Y is X = 3.56 – 0.28.Y -------- (2)

From equation (2) when Y = 3 then X = 3.56 – 0.28.(3) = 2.72

3) From the following data obtain regression equation between height and weight of the F.Y. B.B.A. students. Estimate weight of the student whose height is 160 cm.

Height (in cms)	165	174	170	162	166	165	168	155	150	180
Weight (in kgs)	64	70	66	65	69	63	66	58	55	73

Solution:

Here n = 10

Height X	**Weight Y**	**U**	**V**	**U^2**	**V^2**	**UV**
165	64	0	-1	0	1	0
174	70	9	5	81	25	45
170	66	5	1	25	1	5
162	B = 65	-3	0	9	0	0
166	69	1	4	1	16	4
A = 165	63	0	-2	0	4	0
168	66	3	1	9	1	3
155	58	-10	-7	100	49	70
150	55	-15	-10	225	100	150
180	73	15	8	225	64	120
$\sum X$ = 1655	**$\sum Y$ = 649**	**$\sum U$ = 5**	**$\sum V$ = - 1**	**$\sum U^2$ = 675**	**$\sum V^2$ = 261**	**$\sum UV$ = 397**

➢ Equation of Regression Line Y on X

$$Y = a + b_{yx} . X$$

If u = x – A and v = y – B (where A and B are assumed mean) then

$$b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - (\Sigma u)^2}$$

$$\therefore b_{yx} = \frac{10(397)-(5)(-1)}{10(675)-(-1)^2} = 0.59$$

Also,

$$a = \bar{y} - b_{yx}.\bar{x}$$

Where $\bar{x} = A + \frac{\sum u}{n} = 165 + \frac{5}{10} = 165.5$ and

$$\bar{y} = B + \frac{\sum v}{n} = 65 + \frac{(-1)}{10} = 64.9$$

$$\therefore a = \bar{y} - b_{yx}.\bar{x} = 64.9 - (0.59)(165.5) = -32.745$$

So, regression equation Y on X is Y = - 32.745 + 0.59 X -------- (1)

From equation (1) when Height (X) = 160 then Y = - 32.745 + 0.59 (160) = 61.655 kgs.

4) The following information is obtained from result of an examination:

	Marks in Mathematics (X)	**Marks in Statistics (Y)**
Average	39.5	47.5
S.D.	10.8	16.8
Correlation coefficient between X and Y is 0.42		

Obtain both the regression lines.

Solution:

Equation of Regression Line Y on X

$Y = a + b_{yx} . X$

Where,

$$b_{yx} = r\frac{s_y}{s_x} = 0.42\frac{(16.8)}{(10.8)} = 0.65 \text{ (Approx.)}$$

Also,

$$a = \bar{y} - b_{yx}.\bar{x} = 47.5 - 0.65(39.5) = 21.825$$

So, regression equation is Y = 21.825 + 0.65 X

Now,

Equation of Regression Line X on Y

$X = a' + b_{xy} . Y$

Where,

$$b_{xy} = r\frac{s_x}{s_y} = 0.42\frac{(10.8)}{(16.8)} = 0.27$$

Also,

$$a' = \bar{x} - b_{xy}.\bar{y} = 39.5 - 0.27(47.5) = 26.675$$

So, regression equation is X = 26.675 + 0.27 Y

5) The two regression lines are X + 2Y = 5 and 2X + 3Y = 8 and variance of X is 12. Then find following:

   i. Average of series X and Y.
   ii. Coefficient of Correlation
   iii. Variance of series Y.

Solution:

i. By solving the given regression equation average of X and Y series will be

   X + 2Y = 5 ---------------- (1)

   2X + 3Y = 8 -------------- (2)

   So, after multiplying first regression line by 2 we can eliminate the value of X and Y as follows

   2X + 4Y = 10

   2X + 3Y = 8

Y = 2 So, the average of Y is also 2.

From equation (1) X = 1 So, the average of X is also 1.

ii. Coefficient of correlation is,

Suppose equation (1) is regression equation Y on X.

In that case it can be converted as $Y = \frac{5}{2} - \frac{1}{2}X$ and comparing this equation with $Y = a + b_{yx} . X$.

We can say that $b_{yx} = -\frac{1}{2}$

Now, suppose equation (2) is regression equation X on Y.

In that case it can be converted as $X = \frac{8}{2} - \frac{3}{2}Y$ and comparing this equation with $X = a' + b_{xy} . Y$

We can say that $b_{xy} = -\frac{3}{2}$

Now, $r = \sqrt{b_{yx}b_{xy}}$ and value of r lies between – 1 to + 1.

In our case $r = \sqrt{b_{yx}b_{xy}} = \sqrt{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)} = -0.8660$ which is lies between – 1 to + 1. So, our supposition is correct else we have to change our supposition.

iii. Variance of series Y is

As we know that

$$b_{yx} = r\frac{s_y}{s_x}$$

$$\therefore \left(-\frac{1}{2}\right) = (-0.8660)\frac{S_y}{(3.4641)}$$ (Here, variance of X is 12 so after taking square root S_x = 3.4641)

$$\therefore S_y = 2$$

6) The following table gives the marks obtained by 50 students in Statistics and Mathematics. Find the two regression lines. Also, estimate the marks in Statistics of a student who secured 20 marks in Mathematics.

Marks in Mathematics	**Marks in Statistics**			**Total**
	20 – 25	**25 – 30**	**30 – 35**	
16 – 20	9	14	-	23
20 – 24	6	11	3	20
24 – 28	-	-	7	7
Total	15	25	10	50

Solution:

For bivariate data following calculation is required

Mathematics (y)	**Statistics (x)**								
	20 – 25	**25 – 30**	**30 – 35**	**fy = fv**	**M.V. of y**	**v**	**vfv**	**v^2fv**	**fuv**
16 – 20	(9) 9	(0) 14	-	**23**	18	-1	-23	23	9
20 – 24	(0) 6	(0) 11	(0) 3	**20**	22	0	0	0	0
24 – 28	-	-	(7) 7	**7**	26	1	7	7	7
fx = fu	**15**	**25**	**10**	**n = 50**			∑vfv = -16	∑v^2fv = 30	∑fuv = 16
M.V. of x	22.5	27.5	32.5						
u	- 1	0	1						
ufu	-15	0	10	∑ ufu = -5					
u^2fu	15	0	10	∑ u^2fu = 25					
fuv	9	0	7	∑ fuv = 16					

Where,

$$u=\frac{x-A}{c_x}=\frac{x-27.5}{5} \text{ and } v=\frac{y-B}{c_y}=\frac{y-20}{4}$$

Where

$$\bar{x} = A + \frac{\sum ufu}{n} \times c_x = 27.5 + \frac{-5}{50} \times 5 = 27$$

and

$$\bar{y} = B + \frac{\sum vfv}{n} \times c_y = 22 + \frac{-16}{50} \times 4 = 20.72$$

- Equation of Regression Line Y on X

$Y = a + b_{yx} . X$

Where,

$$\therefore b_{yx} = \frac{n\sum fuv - \sum ufu \sum vfv}{n\sum u^2 fu - (\Sigma ufu)^2} \times \frac{c_y}{c_x}$$

$$\therefore b_{yx} = \frac{50(16) - (-5)(-16)}{50(25) - (-5)^2} \times \frac{4}{5} = 0.47$$

$$a = \bar{y} - b_{yx} . \bar{x}$$

$$\therefore a = \bar{y} - b_{yx} . \bar{x} = 20.72 - 0.47(27) = 8.03$$

So, regression line Y on X is Y = 8.03 + 0.47 X

- Equation of Regression Line X on Y

$X = a' + b_{xy} . Y$

Where,

$$\therefore b_{xy} = \frac{n\sum fuv - \sum ufu \sum vfv}{n\sum v^2 fv - (\Sigma vfv)^2} \times \frac{c_x}{c_y}$$

$$\therefore b_{xy} = \frac{50(16) - (-5)(-16)}{50(30) - (-16)^2} \times \frac{5}{4} = 0.72$$

$$a' = \bar{x} - b_{xy} . \bar{y}$$

$$\therefore a' = \bar{x} - b_{xy} . \bar{y} = 27 - 0.72(20.72) = 12.08$$

So, regression line X on Y is X = 12.08 + 0.72 Y

Now when Y = 20 then X = 12.08 + 0.72(20) = 26.48

❖ Exercise

✓ Theoretical Questions

1) What is Regression? Explain uses of regression in business applications.
2) Give in detail the utility of Regression.
3) Give the difference between Correlation and Regression.
4) State the properties of regression coefficients.

✓ MCQs

1) The independent variable is used to explain the dependent variable in ________.

   a) Linear regression analysis
   b) Multiple regression analysis
   c) Non-linear regression analysis
   d) None of the above

   Answer: a

2) Which of the following statements is true about the regression line?

   a) A regression line is also known as the line of the average relationship
   b) A regression line is also known as the estimating equation
   c) A regression line is also known as the prediction equation
   d) All of the above

   Answer: d

3) Which of the following statements is true about the correlational analysis between two sets of data?

   a) The correlational analysis between two sets of data is known as a simple correlation
   b) The correlational analysis between two sets of data is known as multiple correlation
   c) The correlational analysis between two sets of data is known as partial correlation
   d) None of the above

   Answer: a

4) The coefficient of correlation is the _______ of coefficient of regression.

   a) Reciprocal of product

b) Arithmetic Mean

c) Geometric Mean

d) None of the above

Answer: c

5) Regression analysis is in concerns with prediction of

a) Independent Variable

b) Dependent Variable

c) Constant Term

d) None of the above

Answer: b

6) If r = 0.8, b_{xy} = 0.32 then what will be the value of b_{yx}.

a) 0.48

b) 0.52

c) 2

d) 1

Answer: c

7) If coefficient of correlation r_{xy} = 1, then

a) Regression line become identical

b) Perfect linear co – relationship is observed

c) $b_{yx} = \dfrac{1}{b_{xy}}$

d) All of the above

Answer: d

8) If X + 2Y + 1 = 0 and 2X + 3Y + 4 = 0 are two lines of regression computed from some bivariate data. Then value of coefficient of correlation is

a) – 0.87

b) – 1.15

c) 1.15

d) 0.87

Answer: a

9) If X = 2Y + 4 and Y = kX + 6 are the lines of regression of X on Y and Y on X respectively, find the value of k, if value of r is 0.5.

a) 1 / 8

b) 1 / 3

c) 1 / 2

d) 1 / 4

Answer: a

10) If X – $\overline{X}$ = 25 and Y – $\overline{Y}$ =120, b_{yx} = 2. Find the value of X when Y = 130.

a) 20

b) 25

c) 45

d) 60

Answer: c

✓ **Practical Examples**

1) From the following data

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	5

i. Fit a regression line Y on X and hence predict Y, when X = 10.

ii. Fit a regression line X on Y and hence predict X, when Y = 2.5.

iii. Calculate Karl Pearson's Coefficient of Correlation.

Answer:

i) Y = 2.9 - 0.3 X, Y = - 0.10

ii) X = 3.56 – 0.28 Y, X = 2.86

iii) r = - 0.29

2) Find both the regression equations from the following data and also find Y when X = 65.

X	57	58	59	59	60	61	62	64
Y	77	78	75	78	82	82	79	81

Answer:

X = 16.945 + 0.545 Y, Y = 38.98 + 0.667 X, Y = 82.335

3) From the following information, obtain the line of regression of consumption (Y) on price (X). Hence give an estimate of consumption when price is 124.

Consumption (Y)	**Price (X)**	**Consumption (Y)**	**Price (X)**
250	96	220	112
200	110	220	112
250	100	200	108
280	90	210	116
300	86	300	86
300	92	250	92

Answer:

Y = 572.74 – 3.24 X, Y = 170.98

4) Calculate coefficient of correlation and lines of regression from the following data.

Sales Revenue (‘000 ₹)	**Advertising Expenditure (‘000 ₹)**			
	5 – 15	**15 – 25**	**25 – 35**	**35 – 45**
75 – 125	4	1	-	-
125 – 175	7	6	2	1
175 – 225	1	3	4	2
225 – 275	1	1	3	4

Answer:

r = 0.60, Y = 118.94 + 2.658 X, X = - 1.45 + 0.134 Y

5) An inquire into 50 families to study the relationship between expenditure on accommodation ₹ X and expenditure on food and entertainment, ₹ Y gave the following results:

$\sum X = 8500$, $\sum Y = 9600$, S.D. of X = 60, S.D. of Y = 20, r = 0.6

Estimate the expenditure of food and entertainment when expenditure on accommodation is ₹ 200.

Answer: 198

6) From the following data of rainfall and production of rice find the most likely production corresponding to rainfall of 40 cm

	Rainfall	**Production (Quintals)**
Mean	35	50
S.D.	5	8

Coefficient of correlation = 0.8

Answer:

Y = 5.2 +1.28 X, production is 56.4 quintals

7) Two random variables X and Y have the following regression lines: 12 X – 15 Y + 99 = 0, 60 X – 27 Y – 321 = 0. If the variance of X is given as 36, calculate

i. The average values of X and Y.

ii. The standard deviation of Y.

iii. The coefficient of correlation between X and Y.

Answer:

i) X – $\overline{X}$ = 13, Y – $\overline{Y}$ = 17

ii) S.D. of Y = 8

iii) r = 0.6

UNIT-12 SAMPLING THEORY

12.1 Introduction

In day-to-day life, we are always going to use the statistical information like production of crops, production of goods, per capital income, index of import and export, index number of shares and stock, etc. All above information or statistical data can be collected either by using census enumeration or sample enumeration.

Under the census enumeration or complete enumeration survey method, data are collected for each and every unit of the "Population" or "Universe", which is complete, set of items of interest in any particular situation. For example: If the average yield of wheat in Saurashtra Region is to be calculated, then figure of yield of wheat would be obtained from each and every farm, of Saurashtra Region. Then total yield of complete Saurashtra Region divided by the no. of farms would give the average yield. In this illustration, the total no. of farms in Saurashtra Region taking crop of wheat is considered as the population for determining the average production of wheat.

12.2 Concept of Sampling

Definition: "Aggregate of the units under study or aggregate of measurement of characteristic pertaining to the units under study is known as population."

In the census enumeration the following points should be consider:

(1) Data are collected from each and every unit of the population.

(2) The result obtained is likely to be more representative, accurate and reliable.

(3) It is an appropriate method of obtaining information on rare events such as area under some crops and yield there of, educational level of people, employment of public, etc.

(4) Data of complete enumeration can be widely used as a basis for various surveys.

However, despite these advantages the census method is not very popularly used in practice. The effort, money and time required for carrying out complete enumeration will generally be extremely large and in many cases cost may be so prohibitive that the idea of collecting information may have to be dropped.

While planning a survey for studying certain types of problem, we find that it is not possible as well as practicable to investigate all the units in the population. In such circumstances some of the units are selected from the population and problem is studied on the basis of investigation of such selected units. Such selected units from the population constitute a sample related to the study.

Definition: "Aggregate of some units selected from the units of the population by some definite method is known as sample."

Thus, Sampling is simply the process of learning about the population on the basis of sample drawn from it. In the sampling technique instead of every unit of the universe only a part of the universe is studied and the conclusions are drawn on that basis for the entire universe.

The Process of Sampling involves 3 elements.

(1) Selecting the Sample.
(2) Collecting the information.
(3) Making the conclusions about the population.

The three elements cannot generally be considered in isolation from each other. Sample selection, data collection and estimation are all inter-related and each has an impact on the other.

It should be noted that a sample is not studied for its own sake. The basic objective of its study is to draw inference about the population. In other words, sampling is a tool which helps us to know the characteristic of the universe or population by examining only a small part of it.

The following are some measure points which differentiate between Census and Sample enumeration.

Census Enumeration	**Sample Enumeration**
(1) All the units under study are examined.	(1) Number of units under study are very small to be examined.
(2) Cost of collection and examination of units are more.	(2) Cost of collection and examination of units are less.
(3) It requires more time.	(3) It requires less time.
(4) As number of units under study	(4) As number of units under study

are large, the accuracy of result cannot be maintained.	are less, the accuracy of result can be maintained.
(5) When the field of inquiry is very large. This method is laborious and difficult.	(5) The sample enumeration is less laborious and relatively easy.
(6) If the nature of the unit under study is destructive, this method is not used.	(6) If the nature of the unit under study is destructive this method is most suitable.
(7) If the population is very large, the number of experts may not be available.	(7) In this method, exports are easily available.
(8) This method is rarely used.	(8) Most of the investigation are carried out by sample enumeration.

12.3 Characteristic of a Good Sample

The sample results are to be more reliable, if it possesses the following characteristic.

(1) Representativeness: A sample should be so selected that it is truly represent the universe. To ensure representative ness the random method of selection should be used. i.e. it should be free from prejudice.

(2) Adequacy: The size of sample should be adequate. According to the principle of probability, larger the size of sample, better the properties of the population are reflected in the sample.

(3) Independence: Selection of units in the sample should be done independently of one another. By independence of selection, we mean that the selection of particular item in one draw has no influence on the probability of selection in any other draw.

(4) Homogeneity: If the population is homogeneous means that there is no basic difference in the nature of the universe, the random sample give best result. But if the population is heterogeneous, it should be stratified in various strata and then sample should be selected from all the strata.

(5) Time: Units under study must be select during the same time period of the survey.

12.4 Basic Statistical Laws

The two popular and well-defined statistical laws which indicate the utility of larger size of the sample for the purpose of reducing sampling errors are:

[1] Law of Statistical Regularity

[2] Law of Inertia of Large Number

[1] Law of Statistical Regularity

It states that a reasonably larger number of items selected at random from a large group of items, will on the average, represent the characteristics of the group.

In the words of the statistician W.I. King, “The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group, are almost sure (on the average) to possess the characteristics of the large group".

In other words, this law explains that if a reasonably large sample is selected at random without bias (i.e., probability sampling), it is almost certain that on an average, the sample so chosen, shall have the same characteristics as those of the parent population from where the units constituting the sample have been drawn. It is on the basis of this theory that the law of statistical theory tells us that a random selection is very likely to give a representative sample.

[2] Law of Inertia of Large Number

It states that large groups or aggregate of data show high degree of stability because there is a greater possibility that the extremes on one side are compensated by the extremes on the other side.

The law of inertia of large number is a corollary to the law of statistical regularity. It emphasizes the fact that large numbers are relatively more stable and more reliable than small ones. In a larger number it is unlikely that the data would move in only one direction. Thus, the greater the size of the sample, the greater will be the compensation or tendency to neutralize one another and consequently more stable would be the result. For example, the birth rate, death rate etc. may vary from place to place, in India but India as a whole country, they will be found somewhat stable over a number of years.

The Law of Statistical Regularity and the Law of Inertia of Large Numbers have great importance in the theory of sampling as the sampling error is reduced to minimum if these laws are correctly applied.

12.5 Objectives of Sampling

The main objectives of the sampling theory are:

(1) To estimate population parameter on the basis of sample statistic.

(2) To set the limits of accuracy and degree of confidence of the estimates of the population parameter computed on the basis of sample statistic.

(3) To test significance about the population characteristic on the basis of sample statistic.

12.6 Methods of Sampling

The various methods of sampling or different sampling design can be grouped fewer than two heads random sampling and non-random sampling. The most important difference between random and non-random sampling is that, pattern of sampling variability can be ascertained in case of random sampling, in non-random sampling there is no way of knowing the pattern of variability in the process.

Some of the important sampling methods used in practice are as under.

[1] Random Sampling Methods

(a) Simple random Sampling

(b) Restricted Random Sampling

It is of three types,

(1) Stratified Sampling

(2) Systematic Sampling

(3) Multi-stage Sampling

[2] Non-Random Sampling Methods

(a) Purposive Sampling

(b) Cluster Sampling

(c) Quota Sampling

(d) Convenience Sampling

(e) Sequential Sampling

12.6.1 Random Sampling Methods

(a)Simple Random Sampling

“If each and every unit of the population is given equal chance to select in the sample, the method of sampling is known as Simple Random Sampling.”

In simple random sampling which items get selected in sample is just a matter of chance i.e. personal bias of investigator does not influence the selection. All units of the samples are independently selected, i.e. the selection of any unit does not depend upon the selection of any other unit.

It should be noted that the word ‘random’ does not mean ‘haphazard’ or ‘hit or miss’- it rather means that the selection process is such that the chance only determines which items are included in the sample.

As pointed by Chou, the sample is ‘simple random sample’ if any of the following is true.

(1)All n items of the sample are selected independently of each other and all N items in the population have the same chance of being included in the sample.

(2)At each selection, all remaining items in the population have the same chance of being drawn. If sampling is made with replacement, each item has a probability 1/N of being drawn at each selection. If sampling is made without replacement, the probability of selection of each item at the first draw is 1/N, at the second draw 1/N-1, at the third draw 1/N-2 and so on.

(3)All the possible samples of a given size *n* are equally likely to be selected.

Methods of Drawing SRS:

To ensure randomness of selection one may adopt either Lottery Method or Random Number Table.

(1) Lottery Method: It is the simplest, most common and important method of obtaining a random sample. Under this method all the members of the population or universe are serially numbered on small slips of a paper. They are put in a drum and thoroughly mixed by vibrating the drum. After mixing, the numbered slips are drawn out of the drum one by one according to the size of the sample. The number of slips so drawn constitutes a random sample.

The method would be quite clear with the help of an example. If we want to take sample of 20 items out of population of 500 items, the procedure is to assign the number from 1 to 500 to all the items. Then make a chit/slip of equal size, shape & colour by giving number 1 to 500. Then fold all the slips and put in drum, mixed it. Then make a blindfold selection of 20 slips. The number on slips, that item should be selected as sample from the population.

This method is very popular in lottery draws where a decision about prizes is to be made.

(2) Table of Random Number: The lottery method discussed above becomes quite tedious when the size of population increases. An alternative method of random selection is that of using the table of random numbers.

In this method, sampling is conducted on the basis of random numbers which are available from the random number tables. The various random number tables available are

(1)Tippet’s (1927) random number table.

(2)Fisher and Yates (1938) random number table.

(3)Kendall and Smith (1939) random number table.

(4)Rand Corporation (1955) random number table.

(5)C. R. Rao, Mitra and Mathai (1966) random number table.

Tippets’ table of random number is most popularly used in practice.

1545	7483	2952	7969	9025	5911	3170	8776
3408	5246	2754	1396	6107	8125	6913	6446
1112	9143	0560	2762	2762	5336	4233	7691
1405	6641	2370	6107	1396	1089	8816	2693
3922	9524	4167	9025	9792	6111	1300	5624

➢ **Merits of SRS:**

(1) As all the items of the population have equal chance of being selected in sample, there is no possibility of personal bias or prejudice affecting the result.

(2) As the size of sample increases, becomes increasingly representative of the population.

(3) As compared to judgment sampling a random sample represent the universe in a better way.

(4) The analyst can easily assess the accuracy of this estimate because sampling error follow the principle of chance.

(5) This method is best among all when the population is homogeneous.

➢ **Limitation of SRS:**

(1) In this method a list of all the units of population is required. In absence of such a list, it is difficult for the investigator to use this method.

(2) This method is not proper when the population is heterogeneous.

(3) When the size of population is very small, this method fails to give reliable result.

(4) When the population size is very large, the work of preparing slips or giving number is tedious and tiresome.

(5) Random sampling may produce the most non-random looking result.

(b) Restricted Random Sampling:

(1) Stratified Random Sampling: Stratified Random Sampling Method is one of the random methods which, by using the available information concerning population, attempts to design a more efficient sample than obtained by simple random procedure. If the population is not homogeneous i.e. there are more variation among the units of population, simple random sample can not be representative sample. When this method of sampling is adopted, the population is divided into different group or classes in such a way that (i) there is as great homogeneity as possible within each group and (ii) there is as great heterogeneity as possible between groups. Divided groups or classes called strata and a sample is drawn from each stratum at random. The procedure of making strata is called stratification.

For Example: If we are interested in studying the income pattern of the people of Bombay. The city of Bombay may be divided into various parts such as zone or wards. Here each part is called stratum and from each stratum sample may be taken at random. Before deciding on stratification, we must have knowledge of the complete population. Such knowledge may be based upon expert Judgment, past data, preliminary observation etc.

A stratified sample may be either proportional or disproportionate. In a proportional stratified sampling plan, the number of items drawn from each stratum is proportional to the size of the data. In disproportionate stratified sampling, an equal number of cases are taken from each stratum regardless of how the stratum is represented in the population.

- **Merits of Stratified Random Sampling:**

(1) It is a representative sample of the heterogeneous population.
(2) It lessens the possibility of bias of one sidedness.
(3) Stratified random sampling ensures greater accuracy. The accuracy is maximum if each stratum so formed that it consists of homogeneous items.
(4) As compared to simple random sample, stratified random samples can be more concentrated geographically, i.e. the units from different strata may be selected in such a way that all of them are localized in one geographical area.
(5) If different standards of accuracy are required from different strata, this method is more convenient.
(6) Administrative convenience increases in this type of sampling. By appointing suitable person, accurate information can be obtained from all the parts of the strata or universe.

- **Limitation of Stratified Random Sampling:**

(1) It may be difficult to divide the population into heterogeneous groups.

(2) There may be over-lapping of different strata of the population which will provide an unrepresentative sample.

(3) In the absence of skilled sampling supervisors, the collection of data, stratification procedure and analysis of data is not reliable.

(2) Systematic Sampling: Systematic Sampling method is one of the random sampling methods. This method practically used in those cases where complete list of the population from which sample is drawn is available. This list may be in the form of alphabetical order, geographical order, numerical order or some other order.

A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until the sample has been formed. If we want to select sample of size n from the population of size N, then first of all find sampling interval k as k = n / N.

Then select one item as sample from sampling interval k i.e. from 1 to k, by using lottery method or random number table. Let number of selected item is i ($0 < i \leq k$) then select the subsequent sample at every k^{th} interval, that is i, i + k, i + 2k, i + 3k,........

For Example: In a class there are 150 students with roll number from 1 to 150. It is desired to take a sample of 15 students. By using the systematic sampling select the 15 students.

Solution:

First step in systematic sampling is to find out the sampling interval k.

i.e. $k = \frac{N}{n}$

Here, N = 150, n = 15

$k = \frac{150}{15}$, k = 10

Then select one student at random from roll number 1 to 10. (i.e. 1 to k) suppose the first student comes out to be roll number 8. Then remaining 9 students will be 8 + 10, 8 + 20, 8 + 30, 8 + 40, 8 + 50, 8 + 60, 8 + 70, 8 + 80, ..., 8 + 140.

That is sample should be consist of the roll numbers.

8, 18, 28, 38, 48, 58, 68, 78, 88, 98, 108, 118, 128, 138, 148

- **Merits of Systematic Sampling**

(1) The design of the systematic sampling is very simple and it is convenient to adopt.

(2) The time and work involved in sampling by this method are relatively less.

(3) The result obtained by this method is satisfactory provided care is taken to see that there are no periodic features associated with the sampling interval.

(4)When the population is large the result obtained by this method may be similar to those obtained by proportional stratified sampling.

- **Limitation of Systematic Sampling**

(1) If the population having any hidden periodicities, then the sample obtained by this method are not true representative.

(2) If the list of population units is not prepared in random manner, the expected result obtained for population is not reliable.

(3) Multistage Sampling: In this sampling method, sample of elementary units is selected in stages. Firstly, a sample of cluster is selected and from them a sample of elementary units is selected. It is suitable in those cases where population size is very big and it contains a large number of units.

12.6.2 Non – Random Sampling Method

A sample of elementary units that is being selected on the basis of personal judgment is called a **non – random sampling.** It is of five types.

(a) Purposive Sampling
(b) Cluster Sampling
(c) Quota Sampling
(d) Convenience Sampling
(e) Sequential Sampling

(a) Purposive Sampling: Purposive sampling is the method of sampling by which a sample is drawn from a population based entirely on the personal judgment of the investigator. It is also known as Judgment Sampling or Deliberate Sampling. Randomness finds no place in it and so the sample drawn under this method cannot be subjected to mathematical concepts used in computing sampling error.

(b) Cluster Sampling: Cluster Sampling involves arranging elementary items in a population into heterogeneous subgroups that are representative of the overall population. One such group constitutes a sample for study.

(c) Quota Sampling: In quota sampling method, quotas are fixed according to the basic parameters of the population determined earlier and each field investigator is assigned with quotas of number of elementary units to be interviewed.

(d) Convenience Sampling: In convenience sampling, a sample is obtained by selecting convenient population elements from the population.

(e) Sequential Sampling: In sequential sampling a number of samples lots are drawn one after another from the population depending on the results of the earlier samples drawn from the same population. Sequential sampling is very useful in Statistical Quality Control. If the first sample is acceptable, then no further sample is drawn. On the other hand, if the initial lot is completely unacceptable, it is rejected straightway. But if the initial lot is of doubtful and marginal character falling in the band lying between the acceptance and rejection limits, a second sample is drawn and if need be, a third sample of bigger size may be drawn in order to arrive at a decision on the final acceptance or rejection of the lot. Such sampling can be based on any of the random or non-random method of selection.

12.7 Advantages of Sampling

The sampling techniques have the following advantages over the complete enumeration:

(1) **Less time consuming:** As small numbers of units are to be examined, the survey work as well as analysis work can be completed in less time.

(2) **Less Costly:** In sampling, we study only a part of the population and the total expense of collecting data is less than that required in the census method.

(3) **More reliable result:** The result obtained is generally more reliable than that obtained from a complete enumeration, due to few units are to be examined in sampling.

(4) **More detailed information:** As few units are to be examined, detailed information can be obtained. Thus, the standard of accuracy increases.

(5) Sampling method is the only method that can be used in certain cases. There are some cases in which the census method is inappropriate and the only practicable approach is provided by the sample method. E.g. in destructive type of testing i.e. during inspection, the unit is required to be destroyed; the sampling method is only the way out.

(6) The sampling method is often used to judge the accuracy of the information obtained on a census basis.

(7) Statistical measures, i.e., parameters based on the population can be estimated and evaluated by sample statistic in terms of certain degree of precision required.

(8) It provides a more accurate method of drawing conclusions about the characteristics of the population as parameters.

(9) It is used to draw the statistical inference.

12.8 Sampling and Non-Sampling Error

The main objective of sampling is to estimate the “parameter” of population on the basis of sample “statistic”.

“The difference between the values of parameter and statistic is called sampling error.”

“The error arising due to drawing inference about the population on the basis of few observations or sample is termed sampling error.”

The sampling error does not exist in the complete enumeration survey. Since the whole population is survey.

12.8.1 Sampling Errors

Sampling errors are of two types

(1)Biased Errors: This sampling error is arising due to faulty process of selection of sample, faulty work during the collection of information and faulty methods of analysis.

(2)Unbiased Errors: These errors arise due to chance of difference between the members of population included in the sample and those not included.

12.8.2 Non-Sampling Errors

“Non-sampling errors are errors that occur in acquiring, recording, tabulating statistical data that can not be ascribed to sampling error. They may arise in either in census or a sample survey.”

At a time of collection, classification, tabulation of data errors may be committed, affecting the final result. Error arising in this manner is known as "non-sampling error". Thus, the data obtained in an investigation by complete enumeration is free from sampling error but not free from non-sampling error.

Non-Sampling error may arise due to the following reasons**:**

(1) Errors committed in data processing operations.

(2) Errors committed due to false presentation of data or due to false tabulated results.

(3) Errors may be accounted due to non-response of sample.

(4) Errors may be due to lack of experienced investigator.

(5) Faulty preparation of questionnaire.

(6) Inappropriate statistical units.

❖ Exercise

1. What is sampling? Explain with suitable example.

2. State the advantages and limitation of sampling.

3. Define simple random sampling? Explain the method of selecting simple random sample.

4. Explain stratified random sampling method with its merits and demerits.

5. What is a sampling? Give its objectives.

6. Explain sampling and non-sampling error.

7. State advantages of taking a sample instead of conducting a census.

8. What is random and non-random sampling? Explain?

9. Explain the various reasons for taking sample. Also explain different techniques of random sampling.

10. What are the characteristics of good sample?

11. What are the advantages of sampling?

Answer in Short

1. What is mean by Sampling?

2. Define Simple Random Sampling.

3. Define Stratified Random Sampling.

4. Define Systematic Random Sampling.

5. Define Random Sampling.

Write short note on

1. Simple Random Sampling.
2. Stratified Random Sampling.
3. Sampling and non-sampling error / statistical error.
4. Systematic sampling.

Explain the difference between

1. Simple random sampling and stratified random sampling.
2. Method of complete enumeration and method of sampling enumeration.
3. Random sampling and non-random sampling.

M.C.Q.

1. Sampling can be described as a statistical procedure
   (a) To infer about the unknown universe from a knowledge of any sample
   (b) To infer about the known universe from a knowledge of a sample drawn from it
   (c) To infer about the unknown universe from a knowledge of a random sample drawn from it
   (d) Both (a) and (b).

2. The Law of Statistical Regularity says that
   (a) Sample drawn from the population under discussion possesses the characteristics of the population
   (b) A large sample drawn at random from the population would posses the characteristics of the population
   (c) A large sample drawn at random from the population would possess the characteristics of the population on an average
   (d) An optimum level of efficiency can be attained at a minimum cost.

3. A sample survey is prone to
   (a) Sampling errors (b) Non-Sampling errors
   (c) Either (a) or (b) (d) Both (a) and (b)

4. Statistical decision about an unknown universe is taken on the basis of
   (a) Sample observations (b) A sampling frame
   (c) Sample survey (d) An imaginary population

5. Random sampling implies
   (a) Haphazard sampling
   (b) Probability sampling
   (c) Systematic sampling
   (d) Sampling with the same probability for each unit.

6. Simple random sampling is very effective if
   (a) The population is not very large
   (b) The population is not much heterogeneous
   (c) The population is partitioned into several sections
   (d) Both (a) and (b)

7. Simple random sampling is
   (a) A probabilistic sampling (b) A non-probabilistic sampling
   (c) A mixed sampling (d) Both (b) and (c).

8. Which sampling adds flexibility to the sampling process?
   (a) Simple random sampling (b) Multistage sampling
   (c) Stratified sampling (d) Systematic sampling

9. Which of the following is not a Random number table?
   (a) Tippet's (b) Fisher and Yate's (c) Laplace (d) Rand Corporation

10. Of the following sampling methods, which is a probability method?
    (a) Simple random sampling (b) Judgement sampling
    (c) Quota sampling (d) Convenience sampling

11. Increasing the sample size has the following effect upon the sampling error?
    (a) It increases the sampling error (b) It reduces the sampling error
    (c) It has no effect on the sampling error (d) All the above

12. Among these, which sampling is based on equal probability?
    (a) Simple random sampling (b) Probability sampling
    (c) Stratified sampling (d) Systematic sampling

13. The probability of selecting an item in probability sampling, from the population is known and is:
    (a) Equal to one (b) Equal to zero (c) Non zero (d) All the above

14. Sample is regarded as a subset of
    (a) Data (b) Set
    (c) Distribution (d) Population

15. Choose the correct option regarding the sampling method?
(a) The sample is the part of population.
(b) It helps in determining sampling error.
(c) Sampling saves money, time and energy
(d) All the above

16. Which of these are the steps in the sampling process?
(a) Choosing the sampling frame.
(b) Defining the target population.
(c) Identifying and selecting the method of sampling.
(d) All the above

17. What do we call a complete and proper survey of a population?
(a) Report (b) Census
(c) Sample (d) None of the above

18. Out of these, which is not a probability sampling?
(a) Simple random sampling (b) Cluster sampling
(c) Quota sampling (d) Stratified sampling

19. Out of the mentioned options, which is not a non-probability sampling?
(a) Cluster sampling (b) Judgement sampling
(c) Quota sampling (d) Convenience sampling

20. When the available population is ………, we use a stratified sample.
(a) Too small (b) too large
(c) Heterogeneous (d) homogeneous

:: Answers::

1. (c)	2. (c)	3. (d)	4. (a)	5. (d)	6. (d)	7. (a)	8. (d)	9. (c)	10. (a)
11. (b)	12. (a)	13. (c)	14. (d)	15. (d)	16. (d)	17. (b)	18. (c)	19. (a)	20. (c)

UNIT-13 INDEX NUMBER

13.1 Introduction

We studied averages to find out central value of a distribution and also studied average of deviation from central value. In the Index Number, we will study the average of changes in a group of related variables over time.

Historically, the first index was constructed in year 1764 to compare the Italian price index (year 1750) with the price level of year 1500. Index number originally developed for measuring the effect of changes in price.

Index numbers have become today one of the most widely used statistical devices and there is hardly any field where they are not used.

Production is rising and / or falling, that imports are increasing or decreasing, the crimes are rising in a particular period compared to the previous period as disclosed by the index number. Indices are used to fill the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies. In fact, they are described as barometer of economic activity, i.e., if one wants to get an idea as to what is happing to an economy, he should look to important indices like index number of the Industrial production, agricultural production, business activity etc.

Definition: “Index numbers are devices for measuring differences in the magnitude of a group related variable.”

13.2 Characteristics of Index Numbers

Characteristic of Index Number are as under:

(1) Index number is a specialized average designed to measure the change in-group of related variable over a period of time.

(2) Index number is relative measure; therefore, it does not have any unit of measurement.

(3) Index number measures the effect of changes over a period of time.

(4) Index number measure the net change in a group of related variables.

(5) Index numbers are geometrically compared.

13.3 Uses of Index Numbers

Index numbers are indispensable tools of economic and business analysis. Its significance can be best appreciated by the following points:

(1) Many of the economic and business policies are guided by index number. It helps in framing suitable policies for the country.

(2) The index number provides some guidepost that one can use in making decision.

(3) Index numbers are most widely used in the evaluation of business and economic condition. There is a large number of other fields also where index number is very useful.

For example: Sociologist may speak of population indices. Physiologist measures I.Q., which are essential index numbers comparing a person's I.Q. Score with that of an average of his or her age.

(4) Index numbers reveal trends and tendencies: Since index numbers are most widely used for measuring changes over a period of time, the time series so formed enable us to study the general trends of phenomena under study.

For example: By examining data of imports for India for the last 8 to 10 years, we can say that our imports are showing an upward tendency. Thus, Index numbers are highly useful in studying the general business condition.

(5) Index numbers are very useful in deflating: Index numbers are highly useful in deflating, i.e., they are used to adjust the original data for price changes or to adjust wages for cost-of-living changes and thus transform nominal wages into real wages.

13.4 Issues in Construction of Index Numbers

Before constructing index numbers, a careful thought must be given to the following problems.

(1) **The purpose of Index***:* At the very outset the purpose of constructing the index must be very clearly decided. What the index is to measure and why? There are not all-purpose indices. Every index is of limited and particular use. Thus, a price index i.e. intended to measure consumer's price must not include wholesale price. Failure to decide clearly the purpose of index would lead to confusion and wastage of time with no fruitful result.

(2) **Selection of Base Period:** Whenever index numbers are constructed, a reference is made to some base period. The base period of an index numbers is the period against, which comparisons are made. It may be 0 year, a month or a day. The index for base period (reference period) is always taken as 100. Though the selection of base period would primarily depend upon the objective of the index, the following points need careful consideration of base period:

(a) The base period should be a normal, i.e. it should be free from abnormalities like earthquakes, famines, booms, depression, wars etc.

(b) The base period should not be too distance in the past. It is desirable to have an index based on a fairly recent period. Since comparison with a familiar set of circumstances are more helpful than a comparison with vague condition.

(c) Fixed based or chain base: While selecting the base a decision is to be made as to whether the base shall remain fixed or not i.e. whether fixed base or chain base index. In the fixed base method, the year or a period of years to which all other prices are related is constant for all times. On the other hand, in the chain base method the prices of a year are linked with those preceding years and not with the fixed year.

Difference between fixed base index number and chain base index are as under:

Fixed Base Index	**Chain Base Index**
(1) Uniformity is maintained due to a fixed base, which make interpretation easier.	(1) Every change is measures with respect to the preceding year, which makes a common person difficult to interprete.
(2) This method is useful for comparision of long term fluctuations.	(2) This method is useful for comparision of short term fluctuation
(3) If the base year is very old, selected from past years, certain details may become obstacle.	(3) The base year is changing every year, unrelated items can be replaced by relevant items.
(4) If information for any one year is missing or not available, still index number can be computed for subsequent yer.	(4) If any immediate value is missing or not available the index number for subsequent years can't be computed.

(3) **Selection of number of items**: The items included in an index should be determined by the purpose for which the index is constructed. Every item cannot be included while constructing index numbers and hence one has to select a sample. A decision must be made on the number of commodities to be included and their quantities.

(4) **Price Quotation:** It is a well-known fact that price of many commodities varies from place to place and even from shop to shop in the same market. In order to ensure uniformity, the manner in which prices are to be quoted must also be decided.

There are two methods of quoting prices:

(i) Money prices (ii) Quantity prices.

In money prices, prices are quoted per unit of commodity for example sugar Rs.600 per quintal. In quantity prices, the prices are quoted per unit of money, for example banana 12 units for Rs.10. A decision must also be made as to whether the wholesale prices or retail prices are required.

(5) **Choice of an Average:** Since index numbers are specialized averages; a decision is to be made as to, which particular average should be used for constructing the index. Medians, mode and harmonic mean are almost never used. Basically, a choice has to be made between Arithmetic mean and Geometric mean.

Theoretically speaking, geometric mean is the best average in the constructing of index numbers because of the following reasons:

(a) In the construction of index number, we are concerned with ratios of relative changes and the geometric mean gives equal weights to equal ratio of change.

(b) Geometric mean is less susceptible to major variations as a result of violent fluctuations in the value of individual items.

(c) Index numbers calculated by using this average are reversible and therefore, base shifting is easily possible.

(d) The geometric mean index always satisfies the time reversal test.

Despite theoretical justifications for favouring geometric mean arithmetic mean is more popular used, because arithmetic mean is simpler to compute.

(6) **Selection of appropriate weights:** The problem of selecting suitable a weight is quite important and the same time quite difficult to decide. There are two methods of assigning weights:

(a) **Implicit:** In implicit weighting, a commodity or its variety is included in the index a number of times. Thus, if wheat is to be given in an index twice as much weight as rice, then two varieties of wheat against one of rice may be included in the series.

(b) **Explicit**: In case of explicit weighting, some outward evidence of importance of the various items in the index is given.

Now the question is whether to choose quantity weights or value weights. Another problem is connection with weights is that of deciding whether the weight shall be fixed or fluctuating.

Above are the problems are created at a time of construction of index number as per the purpose of index number weight can be used.

(7) **Selection of an appropriate formula:** A large number of formulas have been devices for constructing index. The problem very often is that of selecting the most appropriate formula. The choice of the formula would depend not only on the purpose of index but also on data available. Prof. Irving Fisher has suggested that an appropriate index is that, which satisfies time reversal test and Factors reversal test. Theoretically, Fisher's method is considered as "ideal" for constructing index number.

13.5 Limitations of Index Numbers

Index number is very popular and widely used in business as well as in economic but it has some limitations or demerits are as under:

(1) Every index is of limited and particular use.
(2) The period that is selected as base should be proper.
(3) In the construction of index number, commodities, price, average, weighted method used properly. Otherwise, it will not give clear picture of the index value.
(4) If there is an error of fixing the weight or calculating the weight, naturally the result will be wrong.
(5) It is only a single digit figure
(6) It does not give any reasons for fluctuations in data.

13.6 Classification of Index Numbers

Index numbers may be classified in terms of what they measure in Economics and Business. The classifications are as under:

(1) Price Index

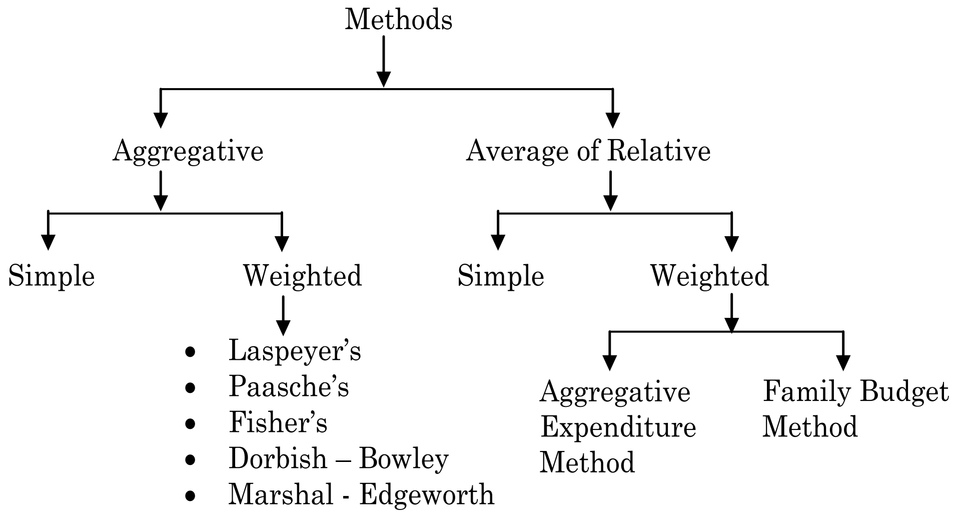
(2) Quantity Index

(3) Value Index

(4) Special Purpose Index number

13.7 Methods of Index Number

Method of construction of Index Number can be shown as under:



13.7.1 Simple Aggregative Method: In this method, the aggregate price of various commodities in a given year is expressed as percentage of the same in the base year. It can be obtained as,

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

where, p_1 = Price of current year

p_0 = Price of base year

Ex.1 Calculate price index by using simple aggregative method from the following data.

Commodity	2023	2024
A	10	14
B	8	10.75
C	0.70	0.95
D	20	23
E	3.60	4.25

Solution:

Here 2023 is considering as base year and 2024 as current year.

Commodity	2023 (p_0)	2024 (p_1)
A	10	14
B	8	10.75
C	0.70	0.95
D	20	23
E	3.60	4.25
Total	$\sum p_0 = 42.3$	$\sum p_1 = 52.95$

Price Index Number for the year 2024 with year 2023 as the base is,

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$P_{01} = \frac{52.95}{42.30} \times 100 = 125.177$$

13.7.2 Weighted Aggregative Method: In this method, appropriate weights are assigned to various commodities to reflect their relative importance in the group. Usually quantities consumed, sold or marketed in the base year or a given year or in some typical years are used as a weight.

If W is weight attached to a commodity, then the price index is obtained as under:

$$P_{01} = \frac{\sum p_1 W}{\sum p_0 W} \times 100 \qquad \text{........(1)}$$

(a) Laspeyres's Index Number:

In this method Laspeyres takes quantity of base year as a weight and calculate Index Number as under:

$$P_{01}{}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \qquad [\text{Put W} = q_0 \text{ in equation (1)}]$$

(b) Paasche's Index Number:

In this method Paasche takes quantity of current year as a weight and calculate Index Number as under:

$$P_{01}{}^{P} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 \qquad [\text{Put W} = q_1 \text{ in equation (1)}]$$

(c) Fisher's Index Number:

In this method, Fisher take Geometric Mean of Laspeyres's and Paasche's Index Number.

That is,

$$P_{01}^{\ F} = \sqrt{P_{01}^{\ L} \times P_{01}^{\ P}} \quad \text{or} \quad P_{01}^{\ F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

Fisher's Index number is called **ideal Index number** due to following reasons:

(i) It is free from bias, since the upward bias of Laspeyer's Index Number is balance to a great extent by the downward bias of Paasche's Index Number.

(ii) It is based on Geometric Mean, theoretically which is considered to be the best average for constructing index number.

(iii) It confirms to certain test of consistency. These tests are Time Reversal Test and Factor Reversal Test.

(iv) This formula takes into account the influence of the current as well as the base year.

(d) Dorbish and Bowley's Index Number:

They suggested the arithmetic mean of the Laspeyres's and Paasche's indices. The formula is

$$P_{01}^{\ DB} = \left(\frac{\dfrac{\sum p_1 q_0}{\sum p_0 q_0} + \dfrac{\sum p_1 q_1}{\sum p_0 q_1}}{2} \right) \times 100$$

$$P_{o1}^{\ DB} = \frac{P_{01}^{\ L} + P_{01}^{\ P}}{2}$$

(e) Marshall and Edgeworth Index Number:

In this method, Marshal and Edgeworth take $W = \frac{q_0 + q_1}{2}$ as general weight in the weighted aggregative formula.

$$P_0^{ME} = \frac{\sum p_1\left(\frac{q_1 + q_0}{2}\right)}{\sum p_0\left(\frac{q_1 + q_0}{2}\right)} \times 100 = \frac{\sum p_1\left(q_0 + q_1\right)}{\sum p_0\left(q_0 + q_1\right)} \times 100$$

$$P_{01}^{ME} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

Ex. 2 The following figures relate to the prices and quantities of certain commodities construct an index number for 2024 with 2021 as base using (i) Laspeyres's (ii) Paasche's (iii) Fisher's (iv) Dorbish and Bowley's (v) Marshal Edgeworth Formula

Commodity	2021		2024	
	Price	Quantity	Price	Quantity
A	28	12	40	10
B	35	8	48	8
C	18	16	32	10
D	20	10	25	12

Solution:

Commodity	2021		2024					
	p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
A	28	12	40	10	336	480	280	400
B	35	8	48	8	280	384	280	384
C	18	16	32	10	288	512	180	320
D	20	10	25	12	250	250	240	300
Total					1104	1626	980	1404

From the above table we have

$$\sum p_0 q_0 = 1104, \sum p_1 q_0 = 1626, \ \sum p_0 q_1 = 980, \ \sum p_1 q_1 = 1404$$

By substituting above values in the formula, we obtained

(i) Laspeyres's Index Number

$$P_{01}{}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{1626}{1104} \times 100$$

$$P_{01}{}^{L} = 147.28$$

(ii) Paasche's Index Number

$$P_{01}{}^{P} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{1404}{980} \times 100$$

$$P_{01}{}^{P} = 143.26$$

(iii) Fisher's Index Number

$$P_{01}{}^{F} = \sqrt{P_{01}{}^{L} \times P_{01}{}^{P}}$$

$$= \sqrt{(147.28)(143.26)}$$

$$P_{01}{}^{F} = 145.25$$

(iv) Dorbish and Bowley Index Number

$$P_{o1}{}^{DB} = \frac{P_{01}{}^{L} + P_{01}{}^{P}}{2}$$

$$= \frac{147.28 + 143.26}{2}$$

$$P_{01}{}^{DB} = 145.27$$

(v) Marshal Edgeworth Index Number

$$P_{01}{}^{ME} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

$$= \frac{1404 + 1626}{980 + 1104} \times 100$$

$$P_{10}^{ME} = 145.39$$

Ex. 3 Calculate Laspeyres's, Paasche's and Fisher's price index for the following data.

Commodity	2022		2023	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution :

Commodity	2022		2023					
	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
Total					200	160	130	103

From the above table we have

$$\sum p_1q_0 = 200,\ \sum p_0q_0 = 160,\ \sum p_1q_1 = 130, \sum p_0q_1 = 103$$

By substituting above values in the formula, we obtained

(i) *Laspeyres's Index Number*

$$P_{01}^{L} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$$

$$= \frac{200}{160} \times 100 = 125$$

(ii) *Paasche's Index Number*

$$P_{01}^{P} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100$$

$$= \frac{130}{103} \times 100 = 126.21$$

(iii) *Fisher's Index Number*

$$P_{01}^{\;F} = \sqrt{P_{01}^{\;L} \times P_{01}^{\;P}}$$

$$= \sqrt{(125)(126.21)} \;\; = 125.6$$

13.7.3 Simple Average of Price Relative: In this method, the prices of each commodity in the current year are expressed as a percentage of the price in the base year.

(i) By using Arithmetic Mean we have,

Price Relative Index Number = $\frac{1}{N}\sum\frac{p_1}{p_0}\times 100$

(ii) By using Geometric Mean we have,

Price Relative Index Number $= AL\left[\frac{\sum \log P}{N}\right]$ $\quad [P = \frac{p_1}{p_0}\times 100]$

Ex. 4 Compute price index from the following data by average of price relative method by using (i) Arithmetic Mean (ii) Geometric Mean.

Commodities	2019 Price	2023 Price
A	200	250
B	300	300
C	100	150
D	250	350
E	400	450
F	500	550

Solution :

Commodity	2019 Price p_0	2023 Price p_1	$P = \frac{p_1}{p_0}\times 100$	Log P
A	200	250	125	2.0969
B	300	300	100	2.0000
C	100	150	150	2.1761
D	250	350	140	2.1461
E	400	450	112.5	2.0511

F	500 550	110	2.0414
Total		737.5	12.5116

(i) Average of Price Relative (Arithmetic Mean)

$$P_{01} = \frac{1}{N}\Sigma\frac{p_1}{p_0}\times 100$$

$$= \frac{1}{6}(737.50)$$

$$= 122.92$$

(ii) Average of Price Relative (Geometric Mean)

$$P_{01} = AL\left(\frac{\Sigma \log P}{N}\right)$$

$$= AL\left(\frac{12.5116}{6}\right)$$

$$= AL\ (2.0853)$$

$$= 121.70$$

13.7.4 Weighted Average of Price Relative: In the weighted aggregative method, price relatives were not computed. However, like simple average of relative method, it is also possible to compute weighted average of relative. For purpose of averaging, we may use either the arithmetic mean or the geometric mean.

(a) By using arithmetic mean we get Index Number as under:

$$P_{01} = \frac{\Sigma PV}{\Sigma V} \qquad \text{where, } P = \frac{p_1}{p_0}\times 100 = \text{Price Relative}$$

$$V = p_0q_0$$

(b) By using geometric mean we get Index Number as under:

$$P_{01} = AL\left[\frac{\Sigma V \log P}{\Sigma V}\right] \qquad \text{where, } P = \frac{p_1}{p_0}\times 100 = \text{Price Relative}$$

$$V = p_0q_0$$

Ex. 5 For the following data compute price index by applying weighted average of price relative method using

(i) Arithmetic mean (ii) Geometric Mean

Commodity	2020		2023
	Price	Quantity	Price
A	3	20	4
B	1.5	40	1.6
C	1	10	1.5

Solution:

(i) Price relative Index Number using weighted arithmetic mean:

Commodity	2020		2023			
	Price	Quantity	Price			
	p_0	q_0	p_1	$V = p_0q_0$	$P = \frac{p_1}{p_0} \times 100$	PV
A	3	20	4	60	133.3	8000
B	1.5	40	1.6	60	106.67	6400
C	1	10	1.5	10	150	1500
Total				$\sum V = 130$	$\sum PV =$	15900

$$P_{01} = \frac{\sum PV}{\sum V}$$

$$P_{01} = \frac{15900}{130} = 122.3$$

(ii) Price relative Index Number using Geometric mean:

Commodity	2020		2023			log P	V log P
	Price	Quantity	Price	V =			
	p_0	q_0	p_1	p_0q_0	$P = \frac{p_1}{p_0} \times 100$		
A	3	20	4	60	133.3	2.1249	127.494
B	1.5	40	1.6	60	106.67	2.0282	121.692
C	1	10	1.5	10	150	2.1761	21.761
Total				130			270.947

$$P_{01} = AL\left[\frac{\sum V \log P}{\sum V}\right]$$

$$= AL\left[\frac{270.947}{130}\right]$$

$= AL\ (2.084)$

$= 121.3$

13.7.5 Construction of Fixed Base Index and Chain Base Index:

The Index Number on a given fixed base was, therefore not affected by changes in the relevant values of any other year. On the other hand, in the chain base method, the value of each period is related with that of the immediately preceding period and not with any fixed period.

For the construction of Index Number by chain base method, using appropriate Index Number formula, a series of Index Number are computed for each year with preceding year as the base year.

$$\text{Chain Base Index} = \frac{\text{Current Year Price}}{\text{Previous Year Price}} \times 100$$

Ex.6 From the following data of wholesale prices of a certain commodities. Construct Index Number by using (i) Fixed Base Method (ii) Chain Base Method

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Price	750	500	650	600	720	700	690	750	840	800

Solution:

$$\text{Fixed Base Index} = \frac{\text{Current Year Pr ice}}{\text{Base Year Pr ice}} \times 100$$

$$\text{Chain Base Index} = \frac{\text{Current Year Pr ice}}{\text{Pr evious Year Pr ice}} \times 100$$

Year	Price	Fixed Base Index	Chain Base Index
2010	750	100	100
2011	500	$\frac{500}{750}\times 100 = 66.67$	$\frac{500}{750}\times 100 = 66.67$
2012	650	$\frac{650}{750}\times 100 = 86.67$	$\frac{650}{500}\times 100 = 130$
2013	600	$\frac{600}{750}\times 100 = 80$	$\frac{600}{650}\times 100 = 92.31$
2014	720	96	120
2015	700	93.33	97.22
2016	690	92	98.57
2017	750	100	108.60
2018	840	112	112
2019	800	106.67	95.24

13.7.6 Construction of Consumer Price Index:

Consumer price index may be constructed by applying any of the following methods.

(1) Aggregate Expenditure Method

(2) Family Budget Method.

(1) Aggregate Expenditure Method:

In this method, quantities of commodities consumed by the particular group in the base year are considered as weight. The prices of commodities for various groups for the current year are multiplied by the quantities of the base year and also with the price of base year.

Aggregate expenditure for current year as well as base year is obtained. Then aggregate expenditure of the current year is divided by the aggregate expenditure of the base year and is multiplied by 100. We obtained Consumer Price Index as under:

$$\text{Consumer price Index } = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

This is the same as Laspeyres's Method discussed earlier.

(2) Family Budget Method:

In this method, first of all the budget of a large number of people for whom the index number is meant are carefully studied and the aggregate expenditure of an average family on various items is estimated. These constitute weights. The weights are thus the value weights obtained by multiplying the prices by quantities consumed i.e. p_0q_0. The price relatives for each commodity are obtained. These price relatives are multiplied by the value weights for each item and this value is divided by the sum of the weights. We obtained Consumer Price Index as under:

$$\text{Consumer Price Index } = \frac{\sum IW}{\sum W} \qquad \text{where, } I = \frac{p_1}{p_0} \times 100 \text{ and W = } p_0q_0$$

$$\text{Consumer Price Index } = \frac{\sum [\frac{p_1}{p_0} \times 100 \times p_0 q_0]}{\sum p_0 q_0}$$

Consumer Price Index $= \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

This is the same as Laspeyres's Method discussed earlier.

Ex. 7 Construct a cost of living/consumer price index from the following data.

Commodity	Index Number	Weight
A	281	46
B	177	10
C	178	7
D	210	12
E	242	25

Solution:

Commodity	Index Number (I)	Weight (W)	IW
A	281	46	12926
B	177	10	1770
C	178	7	1246
D	210	12	2520
E	242	25	6050
Total		100	24512

Consumer Price Index $= \frac{\sum IW}{\sum W}$

$= \frac{24512}{100}$

$= 245.12$

Ex. 8 Construct a cost of living/consumer price index from the following data using (i) Aggregate Expenditure Method (ii) Family Budget Method.

Commodity	2021		2022
	Price	Quantity	Price
A	20	4	24
B	1.25	3	1.50
C	5.00	2	8.00
D	2.00	1	2.25

Solution:

(i) Aggregate Expenditure Method

Commodity	2021		2022		
	Price	Quantity	Price		
	p_0	q_0	p_1	p_0q_0	p_1q_0
A	20	4	24	80	96
B	1.25	3	1.50	3.75	4.5
C	5.00	2	8.00	10	16
D	2.00	1	2.25	2	2.25
Total				95.75	118.75

$$\text{Consumer Price Index} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$$

$$= \frac{118.75}{95.75} \times 100 = 124.02$$

(ii) Family Budget Method :

Commodity	2021		2022			
	Price	Quantity	Price			
	p_0	q_0	p_1	$I = p_0q_0$	$W = \frac{p_1}{p_0} \times 100$	IW
A	20	4	24	80	120	9600
B	1.25	3	1.50	3.75	120	450
C	5.00	2	8.00	10	160	1600
D	2.00	1	2.25	2	112.5	225
Total				95.75		11875

$$\text{Consumer Price Index} = \frac{\sum IW}{\sum W}$$

$$= \frac{11875}{95.75} = 124.02$$

13.8 Test of Adequacy of Index Number Formula

From a statistical point of view, the system of calculation of index number should such that, it satisfies certain mathematical test. These are:

(1) Unit Test
(2) Time reversal test.
(3) Factor reversal test.
(4) Circular test.

(1) Unit Test: This test requires the index number to be independent of the units in which prices and quantities of various commodity are quoted. This test is satisfied by all the formula except simple aggregative.

Since the price relative are suitable for absolute prices either explicitly as in the case of relative methods or implicitly as in the case of aggregative method.

(2) Time Reversal Test: This test implies that if the time subscript of any index formula be interchanged then the resulting index should be reciprocal of the original index,

i.e. $P_{01} \times P_{10} = 1$ (Omitting the factor 100 from each index).

where, P_{01} denotes the index for current period 1 based on base period 0.

P_{10} denotes the index for based period 1 based on current period 0.

This test is satisfied by all the index number except Laspeyer's and Paasche's index.

(3) Factor Reversal Test: This test state that an price index when multiplied by an quantity index with the same base year and average of commodity gives the true value ratio as the product of price and quantity.

P_{01} is price index for given year with reference to the base year.

Q_{01} is quantity index for given year with reference to base year.

Then,

$$P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

(4) Circular Test: This is another test for test of adequacy of an index number. This test was first suggested by Waterford. According to this test, the index should work in circular fashion i.e. if an index number is computed for the period 1 on the base period 0, another index number is computed for 2 on the base period 1 and still another index number is computed for period 2 on the base period 0. And then the product should be equal to 1.

Symbolically, $P_{01} \times P_{12} \times P_{20} = 1$

Ex. 9 For the given data verify the Time Reversal Test and Factor Reversal Test for Laspeyres's, Paasche's and Fisher's Index formula.

Commodity	2018		2022	
	Price	Quantity	Price	Quantity
A	14	28	15	28
B	15	30	16	32
C	13	26	14	27
D	18	25	20	24

Solution :

Time Reversal Test : Time Reversal Test is satisfy if, $P_{01} \times P_{10} = 1$.

Commodity	2018		2022					
	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_0q_1	p_1q_1
A	14	28	15	28	420	392	392	420
B	15	30	16	32	480	450	480	512
C	13	26	14	27	364	338	351	378
D	18	25	20	24	500	450	432	480
Total					1764	1630	1655	1790

Laspeyer's Index Number

$$P_{01}{}^{L} \times P_{10}{}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1}$$

$$= \frac{1764}{1630} \times \frac{1655}{1790} = 0.99$$

Here, $P_{01}{}^{L} \times P_{10}{}^{L} \neq 1$

Therefore, Laspeyer's Index Number does not satisfy the Time Reversal Test.

Paasche's Index Number

$$P_{01}{}^{P} P_{10}{}^{P} = \frac{\Sigma\, p_1 q_1}{\Sigma\, p_0 q_1} \times \frac{\Sigma\, p_0 q_0}{\Sigma\, p_1 q_0}$$

$$= \frac{1790}{1655} \times \frac{1630}{1764} = 0.993$$

$$P_{01}{}^{P} . P_{10}{}^{P} \neq 1$$

Therefore, Paasche's Index Number does not satisfy the Time Reversal Test.

Fisher's Index Number

$$P_{01}{}^{F} . P_{10}{}^{F} = \sqrt{\frac{\Sigma\, p_1 q_1}{\Sigma\, p_0 q_1} \times \frac{\Sigma\, p_1 q_0}{\Sigma\, p_0 q_0} \times \frac{\Sigma\, p_0 q_0}{\Sigma\, p_1 q_0} \times \frac{\Sigma\, p_0 q_1}{\Sigma\, p_1 q_1}}$$

$$= \sqrt{\frac{1790}{1655} \times \frac{1764}{1630} \times \frac{1630}{1764} \times \frac{1655}{1790}} = \sqrt{1} = 1$$

$$P_{01}{}^{F} P_{10F} = 1$$

Therefore, Fisher's Index Number satisfies the Time Reversal Test.

Factor Reversal Test

Factor Reversal Test Satisfy by any Index number formula if,

$$P_{01} Q_{01} = \frac{\Sigma\, p_1 q_1}{\Sigma\, p_0 q_0} = \text{Value Index}$$

Laspeyer's Index Number

$$P_{01}{}^{L} Q_{01}{}^{L} = \frac{\Sigma\, p_1 q_0}{\Sigma\, p_0 q_0} \times \frac{\Sigma\, q_1 p_0}{\Sigma\, q_1 p_1}$$

$$= \frac{1764}{1630} \times \frac{1655}{1790} = 1$$

But, $P_{01}{}^{L}Q_{01}{}^{L} = \dfrac{\Sigma\, p_1q_1}{\Sigma\, p_0q_0}$

$$= \frac{1790}{1630} = 1.015$$

Therefore, Laspeyer's Index Number does not satisfy factor reversal Test

Paasche's Index Number:

$$P_{01}{}^{P}Q_{01}{}^{P} = \frac{\Sigma\, p_1q_1}{\Sigma\, p_0q_1} \times \frac{\Sigma\, q_1p_1}{\Sigma\, q_0p_1}$$

$$= \frac{1790}{1655} \times \frac{1790}{1764} = 1.0915$$

But, $P_{01}{}^{L}Q_{01}{}^{L} = \dfrac{\Sigma\, p_1q_1}{\Sigma\, p_0q_0}$

$$= \frac{1790}{1630} = 1.015$$

Therefore, Paasche's Index Number does not satisfy Factor Reversal Test

Fisher's Index Number

$$P_{01}{}^{F}Q_{01}{}^{F} = \sqrt{\frac{\Sigma\, p_1q_1}{\Sigma\, p_0q_1} \times \frac{\Sigma\, p_1q_0}{\Sigma\, p_0q_0} \times \frac{\Sigma\, q_1p_1}{\Sigma\, q_0p_1} \times \frac{\Sigma\, q_1p_0}{\Sigma\, q_0p_0}}$$

$$= \sqrt{\frac{1790}{1655} \times \frac{1764}{1630} \times \frac{1790}{1630} \times \frac{1655}{1764}}$$

$$= \sqrt{\frac{(1790)^2}{(1630)^2}} = \frac{1790}{1630} = 1.015$$

But, $P_{01}{}^{L}Q_{01}{}^{L} = \dfrac{\Sigma\, p_1q_1}{\Sigma\, p_0q_0}$

$$= \frac{1790}{1630} = 1.015$$

Hence, Fisher's Index Number Satisfies Factor Reversal Test.

:: Exercise ::

1. Define Index Number; State the characteristic of Index Number.

2. State the uses of Index Number.

3. Show that Dorbish and Bowley's index Number does not satisfy time reversal test and factor reversal test.

4. Explain usefulness of Index Number in Economics.

5. State 4 limitation and uses of Index Number.

6. Describe in brief how Index Number is constructed.

7. Discuss the methods of construction, uses and limitation of Index Number.

8. Discuss the various points to be considered in the selection of base year, weights and commodities in the construction of an Index Number.

9. Explain the Time Reversal and Factor Reversal Test of an Index Number. Prove that the Fisher's Index Number satisfies these tests.

10. Why Fisher's Index Number is called ideal? Show that it satisfies both, the time reversal Test as well as Factor Reversal Test.

12. Write short note on

(a) Index number **(b)** Fisher's Index

(c) Construction of Index Number

13. Write answer in short:

(a) Define index number.

(b) What is base year?

(c) Define chain base index number.

(d) What is implicit weight?

(e) Name the important tests of index number.

14. Differentiate between

(a) Fixed base index and chain base index

(b) Simple and Weighted Index Number

Practical example

Ex.1 Calculate price index by using simple aggregative method from the following data.

Commodity	2022	2023
A	100	140
B	80	107.5
C	7	9.5
D	200	230
E	36	42.5

Ex. 2 The following figures relate to the prices and quantities of certain commodities construct an index number for 2024 with 2021 as base using (i) Laspeyres's (ii) Paasche's (iii) Fisher's (iv) Dorbish and Bowley's (v) Marshal Edgeworth Formula

Commodity	2021		2024	
	Price	Quantity	Price	Quantity
A	15	5	22	8
B	5	10	7	15
C	60	1.2	75	2
D	14.25	15	15	25
E	32	18	36	30
F	2.5	8	3	10

Ex. 3 Compute price index from the following data by Average of price relative method by using Arithmetic Mean.

Commodities	2021	2022
	Price	Price
A	25	28
B	30	35
C	375	380
D	36	40
E	440	500
F	265	300

Ex. 4 For the following data compute price index by applying weighted average of price relative method using weighted arithmetic mean Arithmetic mean.

Commodity	2020		2023
	Price	Quantity	Price
A	160	40	200
B	400	25	600
C	50	5	70
D	10	20	18
E	2	10	3

Ex.5 From the following data of wholesale prices of a certain commodities. Construct Index Number by using (i) Fixed Base Method (ii) Chain Base Method.

Year	2012	2013	2014	2015	2016	2017	2018	2019
Price	230	250	230	250	270	280	300	300

Ex. 6 Construct a cost of living/consumer price index from the following data.

Commodity	Index Number	Weight
A	230	48
B	225	18
C	220	8
D	200	12
E	235	14

Ex. 7 Construct a cost of living/consumer price index from the following data using
(i) Aggregate Expenditure Method
(ii) Family Budget Method.

Commodity	2023		2024
	Price	Quantity	Price
A	16	35	18
B	40	25	45
C	60	20	120
D	80	10	90
E	30	20	45
F	28	15	35

Ex. 8 For the given data verify the Time Reversal Test and Factor Reversal Test for Laspeyres's, Paasche's and Fisher's Index formula.

Commodity	2020		2022	
	Price	Quantity	Price	Quantity
A	40	1.5	39	1
B	44	10	40	12
C	50	1.5	45	2
D	36	1.5	30	2

:: ANSWERS ::

1. Price Index Number = 125.177 [Hint: $P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$]

2. $\sum p_0 q_0 = 1006.75, \sum p_1 q_0 = 1167, \sum p_0 q_1 = 1656.25, \sum p_1 q_1 = 1916$

(i) Laspeyres's Index Number = 115.92 [Hint: $P_{01}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$]

(ii) Paasche's Index Number = 115.68 [Hint: $P_{01}^{P} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$]

(iii) Fisher's Index Number = 115.8 [Hint: $P_{01}^{F} = \sqrt{P_{01}^{L} \times P_{01}^{P}}$]

(iv) Dorbish Bowley Index Number = 115.8 [Hint: $P_{o1}^{DB} = \frac{P_{01}^{L} + P_{01}^{P}}{2}$]

(v) Marshal Edgeworth Index Number = 115.77

[Hint: $P_{01}^{ME} = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$]

3. Average of Price Relative (Arithmetic Mean) = 111.33

[Hint: $P_{01} = \frac{1}{N}\sum \frac{p_1}{p_0} \times 100$]

4. Price relative Index Number using weighted arithmetic mean = 145.5

[Hint: $P_{01} = \frac{\sum PV}{\sum V}$]

5.

Year	2012	2013	2014	2015	2016	2017	2018	2019
Price	230	250	230	250	270	280	300	300
F.B.I.	100	108.7	100	108.7	117.39	121.74	130.43	130.43
C.B.I	100	108.7	92	108.7	108	103.7	107.14	100

6. Consumer Price Index = 225.4 [Hint: Consumer Price Index $= \frac{\sum IW}{\sum W}$]

7. (i) Aggregate Expenditure Method = 141.48

[Hint: Consumer Price Index $= \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$]

(ii) Family Budget Method = 141.48 [Hint: Consumer Price Index $= \frac{\sum IW}{\sum W}$]

8. $\sum p_0q_0$ = 669, $\sum p_1q_0$ = 740, $\sum p_0q_1$ = 571, $\sum p_1q_1$ = 629

Laspeyres's Price Index Number = 110.61 Laspeyres's Quantity Index N. = 85.35

Paasche's Price Index Number = 110.16 Paasche's Quantity Index Number = 85

Fisher's Price Index Number = 110.38 Fisher's Quantity Index No. = 85.17

:: M.C.Q. ::

1. Which average is considered as the ideal average in construction of Index Number?
   (a) Weighted Average (b) Arithmetic Mean
   (c) Geometric Mean (d) None

2. Index numbers are known as a specific type of ____.
   (a) Average (b) Correlation
   (c) Dispersion (d) None

3. Index number is also known as economic ____________.
   (a) Parameter (b) Barometer

(c) Constant (d) None

4. Which of the following is known as the ideal index number?
(a) Fisher’s (b) Paasches
(c) Laspeyres (d) None

5. Weight in Laspeyres price index number is _____.
(a) Quantity during the current year (b) Quantity in the base year
(c) Price during the current year (d) Price in the base year

6. In case the values are of equal importance, then the index number is known as
(a) Weighted (b) Composite
(c) Unweighted (d) Value index

7. Value of fisher’s index, given Laspeyres index = 110, Paasche’s index = 108 is
(a) 100 (b) 108
(c) 109 (d) 110

8. In the consumer price index, the household budget method is also known as
(a) Simple Average (b) Average of weights
(c) The weighted average of relatives (d) All of the above

9. Which one of the following is the use of an index number?
(a) To measure changes in quantity
(b) To measure changes in demand
(c) To measure changes in price
(d) To measure changes in variables over a period of time

10. Which index number gives idea of the standard of living of people?
(a) Cost of living index
(b) Fisher’s index
(c) Wholesale price index
(d) Quantity Index

:: ANSWERS ::

1. (c)	2. (a)	3. (b)	4. (a)	5. (b)	6. (c)	7. (c)	8. (c)	9. (d)	10. (a)

UNIT-14 TIME SERIES ANALYSIS

14.1 Introduction

Numerical facts which are collected at different points of time take the form of a time series. The unit of time may be year, a month, a day, an hour etc. Here time is simply a device that enables one relate all phenomena to a set of common stable reference points. In linear regression, two variables have a cause-and-effect relationship, one of the variables can be used to estimate the other. But in time series we use time as independent variable to estimate some other dependent variable.

One of the most important tasks before business man these days is to make estimate for the future. For example, a person who is engaged with a business interested in finding out his likely profit, sales etc. for the next year or for long term planning after 5 to 10 years. So that he could adjust his production accordingly. For the estimates of future value his first step is to gather information from the past. In this connection, one usually deals with statistical data which are collected, observed or recorded at successive interval of time. Such data are generally referred to as "Time Series."

14.2 Definition

A few definitions of Time Series are as under:

(1) "A set of data depending on time is called a time series."-Kenny and keeping

(2) "A series of values over a period of time is called a time series" P.G. Moore.

(3) "A time series is a set of statistical observation arranged in chronological order" - Morris Hamburg.

A set of Numerical observations of the dependent variable, measured at specific points in time in chronological order, usually at equal intervals in order to determine the relationship of time to such variable is known as Time Series.

Time Series data occur naturally in many application areas. For example:

(i) Economics: Monthly data for unemployment, hospital admissions, etc.

(ii) Finance: Daily exchange rate, a share price, etc.

(iii) Environmental: Daily rainfall, air quality readings.

(iv) Social sciences: Population series, such as birthrates or school enrollments.

(v) Medicine: Blood pressure measurements traced over time for evaluating drugs.

14.3 Utility of Time Series Analysis

[1] **It helps in understanding past behaviour:** of the factor which are responsible for the variations. By observing data over a period of time, one can easily understand what changes have taken place in the past.

[2] **It helps in planning future operations:** The understanding of the past behaviour and projecting the past trends are extremely helpful in predicting the future behaviour.

[3] **Evaluation of performance:** The actual performance can be compared with the expected performance and the cause of variation analysed.

[4] **Comparison:** Different time series are often compared and important conclusions drawn there from.

[5] **Estimation of future operation:** Time series study helps in forecasting and planning future operations.

Time series analysis is extremely useful in practical life. Time series analysis is the basis for understanding past behaviour, evaluating current accomplishment, planning future operations. It is also used for comparing the components of different time series.

The analysis of time series makes time series data of great importance not only in business and economics but also for scientists, sociologist, biologist etc.

14.4 Essential Requirements of Time Series

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

Time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure **consistency and reliability**. An extensive data set ensures you have a **representative sample size** and that analysis can cut through irregular movement (noisy) data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting-predicting future data based on historical data.

14.5 Mathematical Models: Decomposition of Time Series

The classical time series decomposition technique is a method for rearranging a time series into Trend, Seasonality, Cyclical and Irregular (noise) components. This method involves representing the time series as the sum of those four components.

For an analysis of time series, the traditional or classical method is to suppose some type of relationship among the components of a time series. The two relationships often called "Models of Time Series" are as follows:

[1] Multiplicative Model: In traditional or classical time series analysis, it is ordinarily assumed that there is a multiplicative relationship between the four components, that is, it is assumed that any particular valued in a series is the product of factors that can be attributed to the various components.

Symbolically: Y = T x S x C x I

Where, Y = Observed value in Time Series

T = Trend

S = Seasonal Components

C = Cyclical Components

I = Irregular Components

[2] Additive Model: Here we assume that the various components of Time Series are additive and the values are the sum of the four components.

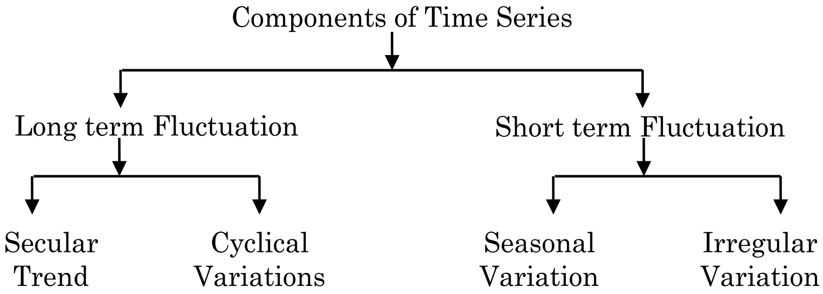
Symbolically: Y = T + S + C + I

14.6 Components of Time Series

The series has revealed that the observations are influenced by a host of causes. Some of these forces are affecting the observations continuously; some operate after equal intervals of time and other are erratic. The forces/causes affecting time series data generate certain movement or fluctuations in a time series. Such characteristic movement or fluctuations of time series are called "Components of Time Series". The components of a time series may be classified into different categories on the basis of operations forces. There are four components of Time Series.

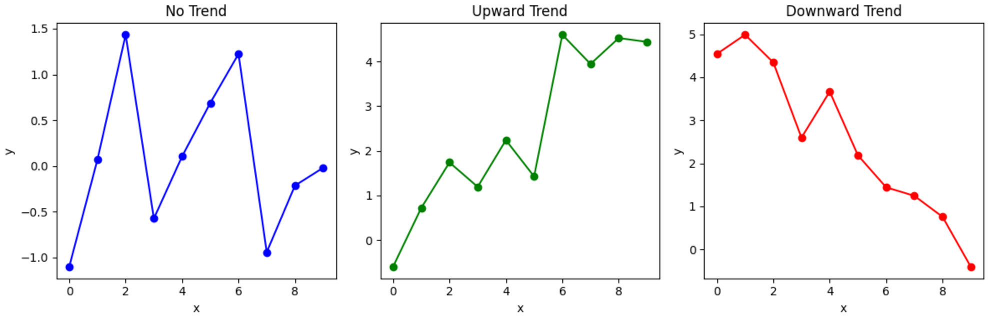
(1) Secular Trend or Trend. (2) Cyclical Variation

(3) Seasonal Variation (4) Irregular Variation



(1) Secular Trend or Trend

The term ‘Trend” is very commonly used in day-to-day conversation. Trend, also called secular or long-term trend, is the basic tendency of production, sales, income, employment etc., to grow or decline over a period of time. In other words, the component of time series which is responsible for its general behaviour over a fairly long period of time as a result of some identifiable influences is called Secular Trend. It is a smooth, regular and long-term tendency of a particular activity to grow or decline. The trend may be upward as well as downward. The concept of Trend does not include Short-range oscillations but rather steady movement over a long time.



It is not necessary that the trend should be in the same direction throughout the given period. We have to observe the general tendency of the data. The time series may be increasing slowly or increasing fast or may be decreasing at various rates or may relatively constant. As long as we can say that the period as a whole characterised by an upward movement or by downward movement. The term “long period of time” is a relative concept which is influenced by the characteristic of the series.

Generally, the longer the period covered, the more significant the trend. When the period is short, the secular movement may be unduly influenced by the cyclical fluctuations. This would make it difficult to separate the various series of variations in time series. Practically to compute the trend, the period must cover at least two to three complete cycles”.

Declining the death rate is an example of downward trend; population growth is an example of upward trend. Demand of cars is the example of upward trend.

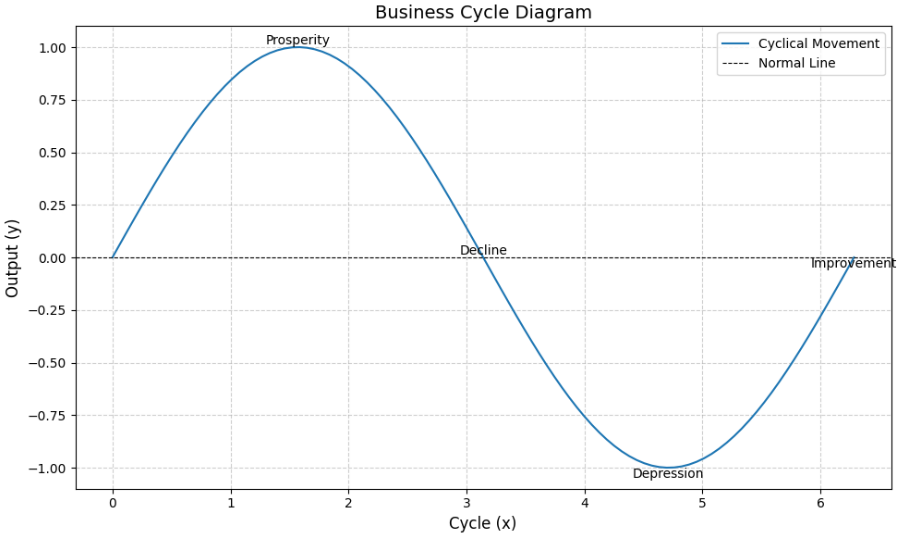
The main objective of measuring the trend is to understand the past behaviour of time series, to forecast the value of time series for long term planning and to study the other components i.e. seasonal, cyclical, irregular by eliminating it.

(2) Cyclical Variation

The oscillatory movement or patterns whose period of oscillation is more than one year are called cyclical variations, one complete period is called a “cycle”. In other words, the term ‘cycle’ refers to the recurrent variation in time series that usually last longer than a year and are regular, neither in amplitude nor in length.

Most of the time series relating to economic and business show some of kind of cyclical variation. There are four well defined period of cycle i.e.

(i) Prosperity (ii) Decline (iii) Depression (iv) Improvement



Each phase changes gradually into the phase which follows it in the order given. The above diagram would illustrate a cycle. In the prosperity phase of the business cycle public is optimistic, prices are high, a period of prosperity is followed by decline until depression is reached. This period for the businessman is pessimistic. Then it follows a period of increasing business activity with rising prices, a period of improvement or recovery. The improvement period generally develops into the prosperity period and a business cycle is completed.

The duration of cycle may vary from 2 to 10 years. But cyclical variation i.e. period cycle does not occur at regular interval and are affected by many erratic, irregular and random forces which cannot be isolated and identified separately.

The purpose of isolating cyclical components in the time series is as follow.

(1) Cyclical measures are useful in formulating policy aimed at stabilizing the level of business activity.

(2) Using the cyclical pattern the businessman can make long term planning for his business.

(3) By studying the cyclical fluctuation in the time series, one can also find the reason for the cyclical movement.

(3) Seasonal Variation

"The component responsible for the regular rise and fall in the time series during a period not more than one year is called seasonal variation." These fluctuations occur in regular sequence and are strictly periodical, period being a year, a month, a week or a day, etc.

Every type of business activity is susceptible to seasonal influenced to a greater or lesser degree and as such these variations are regarded as normal phenomenon recurring every year. Although the word "Seasonal" seems to imply a connection with the season of the year, the term is meant to include any kind of variation which is of periodic in nature and whose repeating cycles are of relatively short duration. Seasonal data are recorded at weekly, monthly or quarterly interval i.e. within year.

The factors that are responsible for seasonal variations are

(i) Natural forces i.e. climate and weather condition:

One of the most important factors causing seasonal variation is climate. Changes in the climate and weather conditions such rainfall, heat, winter, etc. act on different product and industries differently. For example, during summer there is greater demand of cold drinks and water.

(ii) Customs, tradition and habits: Though nature is primarily responsible for seasonal variation in time series, customs and traditions and habits also have their impact. For example: Prices increase during festivals, withdraws from banks are heavy on first week of a month. On certain occasions like Dassehra, Deepawali, Christmas etc., there is great demand of cash money for shopping and gift. Also, in those occasions there is great demand of sweets, rooms on holiday resort etc.

Seasonal variations occur regularly year after year within a fixed period and about the same time and about the same proportion each year. Seasonal variation can occur in both directions upward and downward.

The study and measurement of seasonal pattern constitute a very important part of analysis of a time series. The main purpose of studying the seasonal variations is:

(i) To study the past behaviour of the Time Series.

(ii) To forecast the short time fluctuations.

(iii) Elimination of seasonal variations for measuring cyclical fluctuations.

(4) Irregular Variation

The variations produced by occasional forces, which may operate just once or more than once but without any pattern, system or regularity are called irregular or random or erratic variations. In fact, the category labelled "irregular variation" is really intended to include all types of variations other than those accounting for the trend, seasonal and cyclical variation.

Irregular variations are caused by such isolated special occurrences as flood, earthquakes, strikes, fire, wars, famines, unusual weather, political decisions, and sudden migration. Sudden changes in demand or very rapid technological progress may also be included in this category. By their very nature these variations are very irregular and unpredictable. The irregular influences are confined to very short period of time; they tend to be ironed out over long period of time. Quantitatively it is almost impossible to separate out the irregular variation and cyclical variation. Therefore, while analysing time series the trend and seasonal variations are measured separately and the cyclical and irregular variations are left together.

"To discover and measure any irregularities which characterise the movement of a time series and to isolate them individually is known as the analysis of time series."

:: Exercise ::

Theoretical questions

1. Explain Time Series Analysis.
2. What is the utility of Time Series Analysis?
3. Explain Components of Time series.
4. Explain utility of Time Series Analysis.
5. Explain different mathematical models of Time Series Analysis.
6. What is the primary characteristic of time series data?
7. Provide an example of how time series analysis is used in a real-world application.
8. List and briefly explain the three main components of a time series.
9. Why is it important to decompose time series data?

Write short note on

1. Secular Trend **2.** Seasonal Variation **3.** Cyclical Variation
4. Short term Variation **5.** Decomposition of Time Series

Short questions

1. Define Time Series.
2. Write 3 practical fields where Time Series Analysis used.
3. What are the main essential requirements of Time Series?
4. Give the name of Mathematical Models of Time Series.
5. Define secular trend.
6. What are the phases of cyclical variation?
7. Which factors are responsible for seasonal variation?
8. Due to which reasons, there are irregular variations in time series?

Distinguish between

1. Seasonal components and trend components of a time series.

2. Long term variation and short-term variation in time series.

M.C.Q.

1. Time series is consisting of .... components.
   (a) 1 (b) 2 (c) 3 (d) 4

2. The most important factor causing seasonal variations are
   (a) Change in fashion (b) weather and social customs
   (c) Growth of population (d) technology improvement

3. The period of moving average is to be decided in the light of the length of
   (a) Secular Trend (b) Seasonal Fluctuation
   (c) Cyclical Fluctuation (d) Irregular variation

4. What is the primary characteristic of time series data?
   (a) Cross-sectional data (b) Sequential data (c) Static data (d) Discrete data

5. In which fields can time series analysis be applied?
   (a) Physics and chemistry (b) Economics and finance
   (c) Medicine and healthcare (d) All of the above

6. Which of the following is NOT a component of a time series?
   (a) Trend (b) Seasonal variation (c) Cyclic variation (d) Regular variation

7. What is the purpose of decomposing a time series? (a)
   (a) To make the data easier to understand (b) To remove outliers
   (c) To create new variables (d) To increase data complexity

8. An orderly set of data arranged in accordance with their time of occurrence is called:
   (a) Arithmetic series (b) Harmonic series
   (c) Geometric series (d) Time series

9. A time series consists of:
   (a) Short-term variations (b) Long-term variations
   (c) Irregular variations (d) All of the above

10. The graph of time series is called:
    (a) Non-linear Curve (b) Straight line (c) Histogram (d) Ogive

:: ANSWERS ::

1. (d)	2. (b)	3. (c)	4. (b)	5. (d)	6. (d)	7. (a)	8. (d)	9. (d)	10. (c)

BAOU
Education for All

યુનિવર્સિટી ગીત

સ્વાધ્યાયઃ પરમં તપઃ
સ્વાધ્યાયઃ પરમં તપઃ
સ્વાધ્યાયઃ પરમં તપઃ

શિક્ષણ, સંસ્કૃતિ, સદ્‌ભાવ, દિવ્યબોધનું ધામ
ડૉ. બાબાસાહેબ આંબેડકર ઓપન યુનિવર્સિટી નામ;
સૌને સૌની પાંખ મળે, ને સૌને સૌનું આભ,
દશે દિશામાં સ્મિત વહે હો દશે દિશે શુભ-લાભ.

અભણ રહી અજ્ઞાનના શાને, અંધકારને પીવો ?
કહે બુદ્ધ આંબેડકર કહે, તું થા તારો દીવો;
શારદીય અજવાળા પહોંચ્યાં ગુર્જર ગામે ગામ
ધ્રુવ તારકની જેમ ઝળહળે એકલવ્યની શાન.

સરસ્વતીના મયૂર તમારે ફળિયે આવી ગહેકે
અંધકારને હડસેલીને ઉજાસના ફૂલ મહેંકે;
બંધન નહીં કો સ્થાન સમયના જવું ન ઘરથી દૂર
ઘર આવી મા હરે શારદા દૈન્ય તિમિરના પૂર.

સંસ્કારોની સુગંધ મહેંકે, મન મંદિરને ધામે
સુખની ટપાલ પહોંચે સૌને પોતાને સરનામે;
સમાજ કેરે દરિયે હાંકી શિક્ષણ કેરું વહાણ,
આવો કરીયે આપણ સૌ
ભવ્ય રાષ્ટ્ર નિર્માણ...
દિવ્ય રાષ્ટ્ર નિર્માણ...
ભવ્ય રાષ્ટ્ર નિર્માણ

DR. BABASAHEB AMBEDKAR OPEN UNIVERSITY
(Established by Government of Gujarat)
'Jyotirmay' Parisar,
Sarkhej-Gandhinagar Highway, Chharodi, Ahmedabad-382 481
Website : www.baou.edu.in