BBA 3312
Fundamentals of Statistics
Study Module

স্কুল অব বিজনেস

SCHOOL OF BUSINESS

Bangladesh Open University
বাংলাদেশ উন্মুক্ত বিশ্ববিদ্যালয়

স্কুল অব বিজনেস

SCHOOL OF BUSINESS

বাংলাদেশ উন্মুক্ত বিশ্ববিদ্যালয়

BBA 3312
Fundamentals of Statistics

Course Development Team

Writer

Ariful Islam
Asst. Prof. (Accounting), School of Business
Bangladesh Open University

Editor & Style Editor

Shahima Jabin
Lecturer (Accounting), School of Business
Bangladesh Open University

Coordinator

Dean
School of Business
Bangladesh Open University

This book has been published after being reviewed by the reviewer for the students of School of Business, Bangladesh Open University

This Study Module Book will be used by the BBA students for their study purposes only and it is not for sale.

Print: ....., 2025
Printing by:

CONTENTS

INTRODUCTION TO STATISTICS

1

Unit Highlights

- Introduction
- Definition of Statistics
- The Role of Statistics in Modern Society
- Business Statistics
- Limitation of Statistics

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 1: Introduction to Statistics

Learning Objectives

By the end of this Unit, you should be able to:

1. Clearly explain what statistics is and its significance in various disciplines.
2. Distinguish between descriptive and inferential statistics and their respective applications.
3. Analyze how statistics influence decision-making in modern society, including its impact on business, government, and everyday life.
4. Identify the importance of business statistics and how it is used for market analysis, financial decisions, and operational efficiency.
5. Understand statistical data's limitations, potential biases, and ethical considerations.

1.1 Introduction

Before diving into the formal definition of statistics, it is essential to appreciate its significance and ubiquity in our lives. Statistics is not just a subject studied in academia; it is a fundamental aspect of our daily experiences, influencing decisions in areas as diverse as healthcare, finance, politics, and even our personal lives.

Imagine making everyday decisions, from checking the weather forecast to deciding what to wear to choosing a route based on traffic data—statistics play a crucial role in all these decisions. Businesses use statistics to determine customer preferences, optimize operations, and forecast sales. Scientists use statistical methods to validate hypotheses and interpret experimental data, while governments rely on statistical models to shape policy and allocate resources.

By understanding statistics, we can make informed decisions, interpret information critically, and understand the likelihood of outcomes based on data. Statistics empowers us to consume information, question it, and evaluate its credibility.

Interpreting and making sense of data has become increasingly crucial in an era characterized by data abundance. Statistics is the branch of mathematics that provides the tools and methodologies necessary for this purpose. It transforms raw data into meaningful information, enabling informed decision-making and problem-solving across various fields. Whether in business, healthcare, social sciences, or engineering, statistics is indispensable for understanding and addressing complex issues.

Now, let us define statistics and explore its components and functionalities in the broader context of information analysis and decision-making.

1.2 Definition of Statistics

Statistics is the process of collecting, organizing, analyzing, interpreting, and presenting numerical data. It is a theoretical and applied discipline that deals with studying variability and uncertainty. The word "statistics" derives from the Latin word "status," meaning state or condition. It originally referred to the collection of data about the state. Today, it encompasses many methods and techniques used to understand and make decisions based on data.

Statistics is defined in various ways by different experts and institutions, each highlighting different aspects of the field:

Sir Ronald A. Fisher: "Statistics is the study of populations, the methods of sampling, and the principles and methods of statistical inference." This definition emphasizes the importance of studying entire populations and using samples to make inferences.

Karl Pearson: "Statistics is the grammar of science." Pearson's analogy underscores the foundational role of statistics in structuring scientific knowledge and understanding.

M.G. Bulmer: "Statistics is the science and art of dealing with variability and uncertainty using the collection, analysis, and interpretation of data." Bulmer's view combines the scientific and artistic aspects of handling data.

American Statistical Association (ASA): "Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty." The ASA highlights the dual role of statistics in understanding data and managing uncertainty.

1.3 The Role of Statistics in Modern Society

The role of statistics in modern society is pivotal, as it touches virtually every aspect of human activity. Statistics is omnipresent in contemporary life. From public health decisions to economic policies and educational reforms to environmental conservation efforts, the influence of statistics is profound and far-reaching. By enabling objective data analysis, statistics provides a foundation for informed decision-making across public and private sectors.

Key Areas Influenced by Statistics

1. Government and Policy Making

Governments rely heavily on statistical data to formulate policies, allocate resources, and evaluate the effectiveness of programs. Whether setting economic policy, managing healthcare services, or planning urban development, statistics offer a quantitative basis for decision-making that aims to optimize community outcomes.

2. Healthcare

Statistics are crucial for understanding disease patterns, evaluating treatments, and managing healthcare delivery in the medical field. Statistical analyses drive clinical research, guide public health interventions, and inform decisions on everything from patient care to national health policies.

3. Business and Industry

Businesses use statistics to analyze consumer behavior, optimize operations, and strategize marketing efforts. In industries ranging from finance to manufacturing, statistics help predict market trends, enhance product quality, and maximize efficiency.

4. Science and Technology

Statistics is fundamental to scientific research. It facilitates the design of experiments and the interpretation of results. Statistical analysis is crucial for developing new theories and technologies in fields such as genetics, astronomy, and environmental science.

5. Education

Educational researchers use statistics to assess teaching methods, evaluate learning outcomes, and shape educational policies. By analyzing data on student performance, educators can better understand the effectiveness of curriculum and teaching strategies.

6. Social Sciences

In sociology, psychology, and economics, statistics are indispensable for studying human behavior, social trends, and economic patterns. Statistical methods help researchers uncover relationships between variables and build evidence-based theories.

The Impact on Daily Life

On a more personal level, statistics influence daily decisions and perceptions. Media reports, health information, consumer ratings, and even weather forecasts are all based on statistical data, affecting how individuals process information and make choices.

Ethical Considerations

With the widespread use of statistics comes the responsibility to use data ethically. Misrepresentation or misuse of statistical data can lead to misleading conclusions and potentially harmful outcomes. Thus, ethical data collection, analysis, and reporting practices are paramount to maintaining trust and accuracy.

The role of statistics in modern society is integral and indispensable. As we continue to advance into a data-driven future, the importance of statistics is only set to increase. Understanding and utilizing statistics responsibly and effectively will remain crucial for progress across all aspects of society, from scientific advancements to social well-being.

1.4 Business Statistics

In today's data-driven world, the importance of statistics in business cannot be overstated. Business statistics is vital to modern management, providing the tools and methodologies needed to make informed decisions. It encompasses a wide range of statistical techniques and applications tailored to solving business problems, optimizing operations, and improving overall performance.

The primary objective of business statistics is to extract meaningful insights from data, which can then be used to guide strategic planning, operational efficiencies, and market positioning. It involves a systematic approach to collecting, analyzing, interpreting, and presenting data, ensuring that decisions are based on solid empirical evidence rather than intuition or guesswork.

Business statistics is applicable across various business functions, including marketing, finance, operations, and human resources. For instance, statistics helps in marketing, understanding consumer behavior, segmenting the market, and measuring the effectiveness of marketing campaigns. In finance, it aids in risk management, financial forecasting, and investment analysis. Operations benefit from statistical methods in quality control, process optimization, and supply chain management, while human resources use statistics for performance evaluation, employee satisfaction analysis, and workforce planning.

Key areas of business statistics include descriptive statistics, which focuses on summarizing and describing data, and inferential statistics, which involves making predictions and generalizations about a population based on a sample. Descriptive statistics provide a clear picture of data through measures of central tendency and dispersion and visual representations like charts and graphs. Inferential statistics, on the other hand, allows businesses to make decisions under uncertainty by estimating population parameters, testing hypotheses, and building predictive models.

The role of business statistics extends beyond just analysis and decision-making. It is integral to the entire data lifecycle in a business environment, from initial data collection to the final presentation of findings. Effective use of business statistics requires a strong understanding of statistical principles and the ability to apply these principles in real-world business scenarios.

Moreover, with the advent of big data and advanced analytics, the scope of business statistics has expanded significantly. Modern businesses now have access to vast amounts of data generated from various sources, including digital transactions, social media interactions, and sensor networks. Analyzing this data requires sophisticated statistical techniques and computational tools, making the field of business statistics more dynamic and impactful than ever before.

Therefore, business statistics is an indispensable part of contemporary business practice. It equips managers and decision-makers with the knowledge and skills needed to analyze data effectively, make evidence-based decisions, and drive business success. By leveraging statistical methods, businesses can uncover hidden patterns, predict future trends, and gain a competitive edge in the marketplace.

1.5 Limitation of Statistics

Statistics is a powerful tool for analysis and decision-making, but it also has limitations that must be acknowledged and understood. Recognizing these limitations is crucial for interpreting statistical results correctly and making informed decisions. Here are some key limitations of statistics:

1. Misrepresentation and Misinterpretation

Statistics can be misinterpreted or misrepresented, either intentionally or unintentionally. Charts and graphs can be designed in a way that exaggerates or minimizes certain effects, and statistical results can be taken out of context or based on flawed assumptions, leading to misleading conclusions.

2. Dependence on the Quality of Data

The accuracy and reliability of statistical conclusions heavily depend on the quality of the data used. If data is biased, incomplete, or improperly collected, the results of the statistical analysis will be flawed. This is often summarized by the adage "garbage in, garbage out."

3. Overreliance on Large Data Sets

While large data sets can provide more comprehensive insights, there's also a risk of overfitting models to the data, which means the model is tailored so specifically to the historical data that it may fail to predict future observations accurately. Furthermore, large volumes of data can lead to false discoveries if proper statistical controls (like corrections for multiple comparisons) are not applied.

4. Lack of Causality

Statistics can identify correlations between variables but cannot inherently determine causation. This limitation often leads to misunderstandings where a causal relationship is incorrectly inferred from a mere association. Additional research and analysis are usually required to establish causality.

5. Sensitivity to Outliers

Statistical results can be disproportionately influenced by outliers, which are data points that deviate significantly from other observations. Outliers can skew results and can lead to erroneous interpretations unless they are identified and treated appropriately.

6. Assumptions and Models

Many statistical methods rely on assumptions about the data (such as normality, independence, or homoscedasticity). If these assumptions are not met, the statistical methods may not be appropriate, and the results could be invalid. Choosing the wrong statistical model can also lead to inaccurate conclusions.

7. Complexity and Accessibility

The complexity of statistical techniques can be a barrier to understanding for non-specialists. Misinterpretation of statistical data is common among people who need to understand the methodologies involved thoroughly. This complexity also makes it challenging for general audiences to evaluate the validity of statistical conclusions critically.

8. Ethical Concerns

Statistics can be used to support biased or unethical conclusions, especially when the selection of data or the manner of its analysis is skewed to support a particular viewpoint. Ethical considerations must always be at the forefront when conducting and presenting statistical analyses.

While statistics is indispensable in research and decision-making, its limitations must be carefully considered. Understanding these limitations helps interpret statistical results more critically and responsibly, ensuring that decisions are based on sound statistical reasoning.

Review Questions

1. What is the primary purpose of statistics in data analysis and decision-making?
2. How do descriptive statistics differ from inferential statistics, and can you provide examples of each?
3. Why is statistics considered a crucial tool in modern society, particularly in fields like healthcare and economics?
4. What are the key applications of business statistics in improving organizational performance?
5. Can you describe the role of statistics in formulating public policies and making societal decisions?
6. What are the potential limitations or pitfalls of relying too heavily on statistical analysis?
7. How can statistical data be misinterpreted, and what are some common examples of such misuse?
8. What ethical considerations should be considered when conducting statistical research?

References

Fisher, R. A. (1934). Statistical Methods for Research Workers. *Oliver and Boyd*.

Pearson, K. (1900). The Grammar of Science. *Walter Scott Publishing Co.*

Bulmer, M. G. (1979). Statistical inference. *Principles of statistics*, 165-187.

American Statistical Association Undergraduate Guidelines Workgroup (2014). Curriculum Guidelines For Undergraduate Programs in Statistical Science. *American Statistical Association, Alexandria, VA*. Online: https://www.amstat.org/asa/files/pdfs/EDU guidelines2014-11-15.pdf.

DATA

2

Unit Highlights

- Data
- Data and Information
- Types of Data
- Classifying Data by Level of Measurement
- Sources of Data
- Statistical Procedure
- Data is Collected from either a Population or a Sample
- Example of a Questionnaire for Data Collection

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 2: Data

Learning Objectives

By the end of this Unit, you should be able to:

1. Define what data is and distinguish it from information, understanding their respective roles in analysis.
2. Identify and explain different types of data, including qualitative (categorical) and quantitative (numerical) data, and their significance in statistical analysis.
3. Classify data based on levels of measurement (nominal, ordinal, interval, and ratio) and understand how these classifications impact the choice of statistical methods.
4. Differentiate between primary and secondary data, and describe the characteristics and methods of collecting each type.
5. Understand basic statistical procedures and how data can be collected from either a population or a sample, including practical examples such as creating and using questionnaires for data collection.

2.1 Data

Data refers to raw facts, figures, or information collected for analysis, decision-making, or research purposes. In the context of statistics and data science, data can take many forms, including numbers, words, measurements, observations, or descriptions of things. Data serves as the foundation upon which analysis is built, allowing us to identify patterns, make inferences, and draw conclusions.

Dataset: A structured collection of data, usually presented in a table, where rows represent individual records (cases, observations) and columns represent variables (features, attributes).

Table 2.1 Example of Dataset

Imagine you're managing a small company and have a dataset that records key information about your employees. Here's what the dataset might look like:

Employee ID	Name	Department	Age	Years of Experience	Salary
001	Adyan Isa	Marketing	29	5	Tk. 50,000
002	Inan Khan	Sales	34	8	Tk. 60,000
003	Adib Rahman	IT	27	4	Tk. 55,000
004	Zafir Ahmed	HR	40	15	Tk. 70,000
005	Salma Mehtab	Finance	30	7	Tk. 65,000

Here, we have to understand the Elements, Variables, and Observations

Elements:

Elements are the individual entities or units for which data is collected. In this dataset, each row corresponds to a different employee in the company. Each employee is considered an element.

In this example, the elements are the employees (e.g., Adyan Isa, Inan Khan). Each of these employees is a separate element in the dataset.

Variables:

Variables are the characteristics or attributes that describe the elements. In this dataset, each column (except for the Employee ID) represents a variable that describes an aspect of the employees.

The variables are the characteristics describing the employees (e.g., Name, Department, Age, Years of Experience, Salary).

Like the department in which the employee works (e.g., Marketing, Sales). Like the age of the employee (e.g., 29, 34). Each variable gives us a different piece of information about the employees.

Observations:

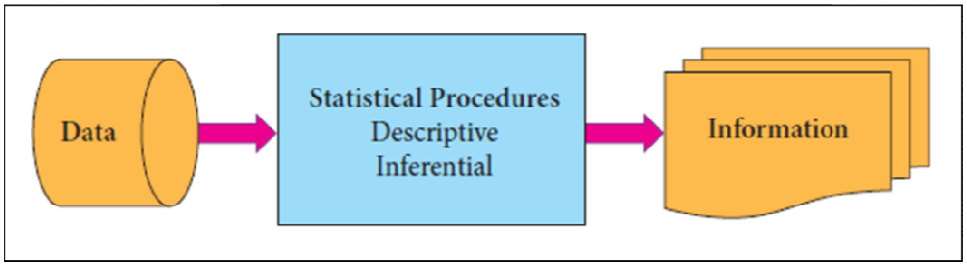
Observations are the specific data points recorded for each element under each variable. In other words, an observation is the value of a variable for a particular element.

For the example of Observations- Adib Rahman's Department = IT, Salma Mehtab's Salary = Tk. 65,000.

2.2 Data and Information?

Data refers to the raw numbers or facts that are collected, such as individual measurements or observations. On its own, data may not provide clear insights.

Information is what you get after processing data using statistical procedures (like descriptive or inferential statistics). This transformed data provides meaningful insights that can be used to make decisions.



Below is a table that outlines the key differences between data and information, based on the definitions provided in the image:

Aspect	Data	Information
Definition	Values assigned to observations or measurements	Data that are transformed into useful facts for specific purposes, such as decision-making
Nature	Raw, unprocessed, and unorganized	Processed, organized, and structured
Purpose	Serves as the raw material or input for analysis	Provides meaning and context for decision-making
Example	A list of test scores or individual sales numbers	Average test score or total sales revenue, which can inform decisions
Transformation	Collected through observations, measurements, or experiments	Derived through statistical procedures (descriptive and inferential)
Usage	Requires further processing to be useful	Directly used for insights, decisions, or understanding
Context Dependency	Independent of context, just a collection of facts	Dependent on context to derive meaning and relevance

2.3 Types of Data

Data in statistics can be broadly categorized into qualitative and quantitative types. Understanding the difference between these types is crucial for selecting the appropriate statistical methods and techniques for analysis.

Imagine you're running a business, and you're trying to understand how well your products are performing in the market. To do this, you collect various pieces of information that will help you make informed decisions. However, as you start gathering data, you realize that not all the information you collect is the same. Some of it is about feelings and opinions, while some of it is hard numbers that you can count or measure. This is where the distinction between qualitative and quantitative data becomes important.

Qualitative Data: Telling the Story Behind the Numbers

Let's say you conduct a customer survey to find out how satisfied people are with your products. You ask questions like, "How satisfied are you with our service?" and "What do you like most about our product?" The responses you get are varied. Some customers might say they are "very satisfied," while others might say they are "neutral" or "dissatisfied." You also receive feedback like "I love the design" or "The product is too expensive."

This type of information is **qualitative data**. It's all about the qualities and characteristics that describe your customers' experiences and opinions. Qualitative data doesn't come in the form of numbers but rather words, categories, or descriptions. It helps you understand the "why" behind customer behavior—why they like or dislike your product, why they might recommend it to a friend, or why they prefer one feature over another. It's like the narrative or story that adds color and context to your business decisions.

Quantitative Data: The Numbers That Drive Decisions

On the other hand, let's consider another scenario where you're looking at your monthly sales report. Here, you're dealing with figures: the number of units sold, the revenue generated, the average price per unit, and the total profit for the month. These are all numbers that can be counted or measured. This is **quantitative data**.

Quantitative data gives you the hard facts—it is the measurable and numerical side of your business. If qualitative data is the story, quantitative data is the plot. It tells you exactly how much you've sold, how much you've earned, and where your sales are strong or weak. You can use this data to calculate trends over time, compare performance between different months, or even forecast future sales. It is precise, objective, and critical for making decisions based on evidence rather than just intuition.

As you look at both types of data, you realize that they complement each other. The qualitative data provides the context and reasons behind the numbers, while the quantitative data gives you the scale and measurement. Together, they offer a full picture of your business performance.

For example, if your quantitative data shows a dip in sales, your qualitative data might reveal that customers are unhappy with a recent change in product design. Conversely, if customer satisfaction scores are high but sales aren't growing, you might explore whether your pricing strategy needs adjustment.

2.4 Classifying Data by Level of Measurement

1. Qualitative (Categorical) Data:

This type of data represents categories or groups. It is non-numeric and is used to describe characteristics or qualities.

Levels of Measurement:

a. Nominal:

Nominal data is the simplest form of data. It categorizes data without a specific order.

Examples: Gender (Male, Female), Eye Color (Blue, Brown, Green), Marital Status (Single, Married, Divorced).

b. Ordinal:

Ordinal data also categorizes data but with a meaningful order or ranking among the categories. However, the intervals between ranks are not necessarily equal.

Examples: Education Level (High School, Bachelor's, Master's), Customer Satisfaction (Satisfied, Neutral, Dissatisfied).

2. Quantitative (Numerical) Data:

a. Discrete Data:

Discrete data consists of distinct, separate values. It arises from counting and usually involves whole numbers. Discrete data cannot take on every possible value within a range; instead, it takes on specific values.

Examples:

Number of students in a class: You can have 25 students, but not 25.5 students.

Number of cars in a parking lot: You can count 50 cars, but not 50.2 cars.

Level of Measurement:

Discrete data can be found in both Interval and Ratio levels of measurement. However, it is more common in Ratio data because many counting processes involve a true zero point (e.g., zero students, zero cars).

b. Continuous Data:

Continuous data can take on any value within a range and is typically measured, not counted. It can include decimals and fractions, representing data that can be divided infinitely.

Examples:

Height of individuals: Height can be 170.5 cm, 172.3 cm, etc.

Time to complete a task: Time can be 3.25 hours, 3.5 hours, etc.

Level of Measurement:

Continuous data can also be found in both Interval and Ratio levels. It is more common in the Ratio level, where there is a true zero (e.g., height, weight), but can also exist in the Interval level (e.g., temperature, which can have negative values but no true zero).

2.5 Sources of Data

Data can be collected from various sources depending on the nature of the study, the type of data needed, and the research objectives. Here's an overview of the main sources of data:

1. Primary Data

Primary data refers to data collected firsthand by the researcher for a specific purpose or study. This data is original and gathered directly from the source, meaning it is specific to the research question at hand.

Characteristics of Primary Data:

Originality: It is collected directly from the source, making it unique and tailored to the researcher's needs.

Specificity: Primary data is usually collected with a particular research question or objective in mind, ensuring its relevance to the study.

Timeliness: It reflects the current situation or recent events, which can be particularly important in dynamic environments.

Methods of Collecting Primary Data:

a. Surveys and Questionnaires

Researchers design a set of questions and distribute them to a target group to gather data on specific topics.

Example: A company surveys customers to understand their satisfaction with a new product.

b. Experiments

Researchers manipulate variables in a controlled environment to observe their effects, collecting data from the outcomes.

Example: A pharmaceutical company conducting clinical trials to test the efficacy of a new drug.

c. Observations

Researchers observe and record behaviors or events as they occur naturally, without intervening.

Example: An ethnographer observing cultural practices in a remote village.

d. Interviews

Researchers engage in direct conversations with individuals to gather detailed information on specific topics.

Example: A journalist interviewing eyewitnesses to gather firsthand accounts of an event.

e. Focus Groups

Researchers bring together a small group of people to discuss specific topics, collecting qualitative data through their interactions.

Example: A marketing team conducting a focus group to explore consumer reactions to a new advertising campaign.

2. Secondary Data

Secondary data refers to data that has already been collected, processed, and published by others. Researchers use secondary data to complement primary data or to inform studies where primary data collection is not feasible.

Characteristics of Secondary Data:

- **Pre-existing:** It has been collected by someone else for a different purpose but can be repurposed for new research.
- **Broad Scope:** Secondary data often covers a wide range of topics, times, and geographic locations, making it useful for trend analysis and comparative studies.
- **Cost-Effective:** Since the data is already available, using secondary data is generally less expensive than collecting primary data.

Sources of Secondary Data:

a. Government Publications
Reports and statistics published by government agencies on topics like demographics, health, and economics.

Example: National census data used to analyze population trends.

b. Academic Journals and Research Papers
Published studies and articles that provide data from various academic and scientific investigations.

Example: A literature review using previous studies on climate change to inform a new research project.

c. Industry Reports
Detailed analyses and statistics published by market research firms, industry groups, or consulting companies.

Example: A report on global smartphone market trends used by a tech company for strategic planning.

d. Historical Records
Documents, archives, and data from past events, are often used for research in fields like history, sociology, and economics.

Example: Historical economic data used to study the Great Depression.

e. Databases
Large, organized collections of data, often available through subscription or purchase, containing information on various topics.

Example: Accessing the *World Bank* database to analyze global economic indicators.

f. Social Media and Online Sources
Data collected from social media platforms, websites, and online forums, is often used for real-time analysis of public opinion or trends.

Example: Analyzing Twitter data to gauge public sentiment on a political issue.

2.6 Statistical Procedure

Descriptive Statistics helps us understand and summarize the characteristics of a specific dataset using numerical and graphical methods. It tells us “What is" in the data without making inferences about a larger population.

Inferential Statistics uses sample data to make predictions, decisions, or generalizations about a larger population. It allows us to go beyond the data at hand and make informed conclusions about a broader context.

2.7 Data is Collected from either a Population or a Sample

Population

A population consists of all the items, individuals, or events about which you want to conclude. It is the complete set that you are interested in studying. In essence, the population is the "large group" that includes every member of the group you are examining.

Examples:

- All the citizens of a country when studying national voting behavior.
- All products manufactured in a factory if you’re assessing quality control.
- All students in a school district when analyzing educational outcomes.

When the population is well-defined, any data collected from it gives a complete picture, but collecting data from an entire population is often impractical due to time, cost, or logistical constraints.

Sample

A sample is a subset of the population selected for analysis. It represents a smaller group taken from the larger population. The sample is the "small group" that is used to make inferences about the entire population.

Examples:

- A group of 1,000 voters selected from across the country to predict the outcome of a national election.
- A few dozen products randomly selected from a day’s production in a factory to check for quality.
- A survey of 200 students from different schools in a district to study overall educational performance.

The sample must be representative of the population to make valid inferences. Proper sampling methods ensure that the sample accurately reflects the characteristics of the population, allowing researchers to make predictions or generalizations about the population based on the sample data.

2.8 Example of a Questionnaire for Data Collection

Expanded Sample Questionnaire: Understanding Lifestyle and Preferences of College Students

Demographics:

1. **Name**:
   (Open-Ended, Text)

2. **Date of Birth**:
   (Open-Ended, Date)
3. **Age**:
   (Calculated from Date of Birth, Ratio Variable)
4. **Gender**:
   - Male
   - Female
   - Non-Binary
   - Prefer not to say
   (Nominal Variable)
5. **Ethnicity**:
   - Asian
   - Caucasian
   - African American
   - Hispanic
   - Other: _____
   (Nominal Variable)

Academic Information:

6. **Major**:
   (Open-Ended, Text, Nominal Variable)
7. **Minor**:
   (Open-Ended, Text, Nominal Variable)
8. **GPA**:
   (Open-Ended, Text, Ratio Variable)
9. **Year in School**:
   - Freshman
   - Sophomore
   - Junior
   - Senior
   (Ordinal Variable)

Lifestyle Choices:

10. **How many hours do you sleep per night?**
    (Open-Ended, Ratio Variable)
11. **How many meals do you eat per day?**
    (Open-Ended, Ratio Variable)
12. **On a scale of 1 to 5, how important is exercise to you?**
    - 1 (Not at all Important)
    - 2 (Slightly Important)
    - 3 (Neutral)
    - 4 (Important)
    - 5 (Very Important)
    (Ordinal Variable)

Social Media Preferences:

13. **Which social media platforms do you use regularly?**
    - ☐ Facebook
    - ☐ Twitter
    - ☐ Instagram
    - ☐ LinkedIn
    - ☐ TikTok
    - ☐ Snapchat

    (Nominal Variable)
14. **On a scale of 1 to 10, how would you rate your happiness with your current social media usage?**

    ☐ _____

    (Interval Variable)
15. **On a scale of 1 to 10, how stressful do you find social media to be?**

    ☐ _____

    (Interval Variable)

Financial Information:

16. **Monthly Income**:
    (Open-Ended, Ratio Variable)
17. **Do you have a budget?**
    - ☐ Yes
    - ☐ No

    (Nominal Variable)
18. **How do you mostly pay for your expenses?**
    - ☐ Savings
    - ☐ Parents
    - ☐ Scholarship
    - ☐ Loan

    (Nominal Variable)

Extended Explanation of Types of Data Variables:

1. **Nominal Variable**: These are categorical variables with no inherent order. Examples include gender, ethnicity, majors, and social media platform preference. You use these when you want to categorize data.
2. **Ordinal Variable**: These have an inherent order, but the distance between categories is not meaningful. Examples include year in school, GPA ranges, and exercise importance ratings. These are useful when the order is meaningful, but you can't make quantitative comparisons like "twice as much" or "three times less."
3. **Interval Variable**: These variables have meaningful distances between categories, but no true zero point. For example, a scale of 1 to 10 for happiness or stress in social media usage allows us to say that an 8 is twice as happy as a 4, but it doesn't mean that a 0 would indicate "no happiness."

4. **Ratio Variable**: These are like interval variables but with a true zero point. Examples include age, GPA, number of hours slept, number of meals, and monthly income. A zero point here is meaningful: zero hours slept means no sleep, zero income means no money earned, etc.

Review Questions

1. What is the difference between data and information, and how are they related?
2. Describe two main types of data and provide examples for each.
3. What are the characteristics of qualitative (categorical) data, and how does it differ from quantitative (numerical) data?
4. Explain the four levels of measurement (nominal, ordinal, interval, and ratio) and provide an example of each.
5. What is the distinction between discrete and continuous data? Give an example of each type.
6. What are primary data sources, and what are some common methods of collecting primary data?
7. List and describe at least three sources of secondary data. How can secondary data be useful in research?
8. What are the advantages and disadvantages of using secondary data compared to primary data?
9. Explain the difference between descriptive and inferential statistics. When would you use each?
10. Define what is meant by a population in statistical analysis. How does it differ from a sample?
11. Why might a researcher choose to use a sample instead of the entire population when conducting a study?
12. What are some key considerations when designing a questionnaire for data collection?

DATA PRESENTATION

3

Unit Highlights

- Purpose of Data Presentation
- Forms of Data Presentation
- Classification of Data
- Tabulation of Data
- Charting Data

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 3: Data Presentation

Learning Objectives

By the end of this Unit, you should be able to:

1. Understand the purpose of data presentation and its importance in statistical analysis and communication.
2. Identify and differentiate between various forms of data presentation, including qualitative and quantitative classification, array formation, and frequency distribution (exclusive and inclusive methods).
3. Organize numerical data effectively through tabulation, including relative and percent frequency distributions.
4. Visualize categorical data using graphical displays such as bar charts, Pareto charts, pie charts, side-by-side bar charts, and doughnut charts.
5. Present numerical data graphically through histograms, cumulative polygons (ogives), and other appropriate methods for clear interpretation and analysis.

3.1 Purpose of Data Presentation:

After data has been collected, it needs to be presented in a form that is easy to understand. Proper data presentation is crucial because raw data, in its unorganized state, can be difficult to comprehend. The goal of data presentation is to transform this raw data into a format that is accessible and informative.

3.2 Forms of Data Presentation:

Data can be presented in two basic forms:

1. **Statistical Tables:**

A statistical table presents numbers in a logical arrangement. It organizes data systematically, with brief explanations to clarify what the numbers represent. Tables are useful for displaying detailed data that can be easily referenced and compared.

2. **Statistical Charts or Graphs:**

A statistical chart or graph is a pictorial device used to present data. Charts and graphs visually represent data, making it easier to see patterns, trends, and relationships within the data. They are particularly effective for communicating complex information quickly and clearly.

For data presentation, the following process should be done.

Classification of Data: Discusses how data is categorized and grouped to facilitate analysis.

Tabulation of Data: Covers the creation and use of tables to organize and display data systematically.

Charting Data: Focuses on the use of various types of charts and graphs to visually present data.

3.3 Classification of Data

Classification of data is a pivotal step in the journey of data analysis. The process takes raw, unorganized data and arranges it into categories or groups based on shared characteristics. This step is not just a formality—it's the foundation upon which meaningful analysis is built,

enabling analysts to identify patterns, relationships, and insights that might otherwise be obscured in a sea of information.

Imagine having a mixed bag of marbles, each differing in color, size, and weight. If you were asked to make sense of this collection, your first instinct would likely be to sort them—perhaps grouping them by color, size, or weight. This act of sorting is akin to data classification. By organizing the marbles into categories, you transform chaos into order, making it easier to conclude, such as which color is most common or which group of marbles is the heaviest. Effective classification lays the groundwork for meaningful interpretation and decision-making in various disciplines, including statistics, machine learning, and data mining (Abdullah, 2019; Aggarwal & Aggarwal, 2015).

3.3.1 Methods of Classification:

There are several ways in which data can be classified. The most common methods include:

I. Qualitative and Quantitative Classification
II. Array Formation
III. Frequency Distribution

I. Qualitative and Quantitative Classification:

In the field of data analysis, classification plays a crucial role in organizing information in a way that makes it more manageable and easier to analyze. Two fundamental methods of classification are qualitative and quantitative classification, each of which categorizes data based on different characteristics.

Qualitative Classification: Qualitative classification deals with data that is categorized based on attributes or qualities that cannot be measured numerically. Instead of focusing on numbers, this method classifies data based on characteristics like gender, hair color, literacy, religion, and other non-numeric attributes. The key aspect of qualitative classification is that the characteristic in question—referred to as an "attribute"—is either present or absent in the population being studied. For example, in a survey that classifies people by hair color, the attribute might be "blond," and the classification would simply indicate whether each individual has blond hair or not. This method is particularly useful when the data pertains to characteristics that are descriptive rather than numerical.

Quantitative Classification: On the other hand, quantitative classification organizes data based on characteristics that can be measured and expressed in numbers. This type of classification is associated with "variables," which are measurable attributes such as height, weight, income, or sales figures. For instance, if you were to classify data based on income levels, the quantitative classification would arrange individuals or entities according to their numerical income values. Quantitative data allows for a wide range of statistical analyses, including comparisons, correlations, and trend analysis, because it deals with numbers that can be mathematically manipulated.

II. Array Formation:

Array Formation is a straightforward yet essential method of data classification. It involves organizing data in a specific order, either ascending (from smallest to largest) or descending (from largest to smallest). This type of classification helps in structuring data so that it's easier to interpret and analyze.

For instance, let's consider the example given in the slide. The income data of ten families is listed as follows:

457, 567, 1004, 1847, 809, 2115, 4391, 918, 5410, and 7348.

These figures are initially unorganized, making it challenging to draw any immediate conclusions about the income distribution among these families.

By applying Array Formation, we can arrange these income values in either ascending or descending order. For example, in ascending order, the data would be organized from the lowest to the highest income:

457, 567, 809, 918, 1004, 1847, 2115, 4391, 5410, 7348

This ordered arrangement allows us to quickly identify the lowest and highest incomes and makes it easier to compute statistical measures such as the median or quartiles.

III. Frequency Distribution:

Frequency distribution is the process of organizing a set of data into categories or intervals and recording the number of observations that fall within each interval. It is a crucial step in data analysis because it allows statisticians to see patterns, trends, and distributions that are not immediately obvious from raw data alone.

Suppose you have the following data representing the scores of 30 students on a test out of 100:

Test Scores:

85, 90, 88, 70, 65, 95, 80, 75, 60, 85, 90, 78, 65, 55, 88, 72, 70, 95, 85, 60, 70, 80, 55, 85, 60, 95, 70, 65, 75, 85

You can organize these scores into a frequency distribution table by grouping them into intervals (bins) and counting how many scores fall into each interval.

Frequency Distribution Table:

Score Range	Frequency
50-59	3
60-69	5
70-79	7
80-89	8
90-99	7

There are different parts in frequency distribution.

1. Range (0-25):

The range is the difference between the maximum and minimum values in a dataset. It indicates the spread of the data across the spectrum.
In this example, maximum score = 95 and minimum score = 55
Therefore, Range = 95 - 55 = 40

2. Class Interval (5 class intervals):

Class intervals divide the range of data into equal segments. Each interval groups a subset of data values. For example, if the range is 55-95, with 5 intervals,
For example, if we decide to divide the scores into 5 intervals:
In the frequency distribution table: Class Intervals are: 50-59, 60-69, 70-79, 80-89, 90-99
Class Interval Width = 59 - 50 + 1 = 10

3. Tally:

A tally is a simple method used in frequency distribution to count and record frequencies of data values using marks or strokes. It is beneficial for organizing raw data before creating a frequency table. Each occurrence of a value is represented by a vertical stroke (|). For every fifth occurrence, a diagonal stroke (/) crosses the previous four strokes, forming a group of five. This grouping makes it easy to count frequencies quickly.

4. Frequency (number of units occurring in each class interval):

Frequency counts the number of data points within each class interval. It helps to see how the data is distributed across the intervals.

5. Class Limit:

Class limits define the boundaries of each class interval. The lower limit is the smallest value in the interval, and the upper limit is the largest value that can belong to that interval.

For the 50-59 interval: lower class limit = 50, upper class limit = 59, and so on for other intervals.

6. Midpoint:

The midpoint of a class interval is calculated by taking the average of the upper- and lower-class limits. It represents the central value of the interval and is useful in various statistical calculations.

It is calculated as: $Midpoint = \frac{\text{Lower Class Limit+Uppe Class Limit}}{2}$

For 50-59 interval: Midpoint = (50 + 59) / 2 = 54.5

For 60-69: Midpoint = (60 + 69) / 2 = 64.5

Here's a summary table combining all the components:

Class Interval	Tally	Frequency	Class Limits	Midpoint
50-59	III	3	50, 59	54.5
60-69	𝍸	5	60, 69	64.5
70-79	𝍸 II	7	70, 79	74.5
80-89	𝍸 III	8	80, 89	84.5
90-99	𝍸 II	7	90, 99	94.5

Determining the Number of Classes and Class Intervals

When organizing data into a frequency distribution, one must decide how many groups, or "classes," to create and how wide each class should be. This process is fundamental for making raw data more interpretable and meaningful.

To begin, the first step involves determining the number of classes. While the number of classes is not fixed, certain guidelines exist to strike a balance between too much detail (which can overwhelm readers) and oversimplification (which can obscure patterns in the data). A widely accepted approach is Sturges' Rule, which provides a formula for calculating the ideal number of classes based on the size of the dataset.

The rule suggests that the number of classes (k) can be estimated using the formula:

k=1+3.322$\log_{10}$N

Here, *N* represents the total number of observations in the dataset, and $\log_{10}$ is the logarithm to base 10. For instance, if a dataset contains 100 observations, the logarithm of 100 is 2. Using the formula, the number of classes would be calculated as follows:

$k=1+3.322\times2=7.644\approx8$

Thus, the data would be divided into eight classes in this example.

The next step is determining the class intervals, which define the range of values each class covers. This requires calculating the range of the data, which is the difference between the largest and smallest values in the dataset. For example, if the maximum value in a dataset is 90 and the minimum value is 10, the range is:

$\text{Range}=90-10=80$

Once the range is determined, the class width can be calculated by dividing the range by the number of classes. For the example above, if there are eight classes, the class width would be:

$$\text{Class Width} = \frac{\text{Range}}{\text{Number of Classes}} = \frac{\text{Range}}{k=1+3.322\log_{10}N} = \frac{80}{8}=10$$

This means each class will cover a span of 10 units. Starting from the minimum value, the first class interval would be 10–20, the second 20–30, and so on, until the maximum value is covered.

Class intervals should be mutually exclusive (no overlaps) and exhaustive (covering all possible values in the dataset) to ensure clarity. They should also be rounded to convenient numbers for ease of understanding. For instance, a class interval of 9.25 units might be rounded up to 10 for simplicity.

It is important to verify that the chosen intervals accommodate all observations and reflect the purpose of the analysis. If the dataset is highly skewed, unequal intervals may be used to represent the data more accurately. For example, in a dataset of incomes, smaller intervals might be used for lower incomes, while larger intervals might be chosen for higher incomes to capture variations effectively.

Importance of Frequency Distribution:

1. **Simplification of Complex Data:** Large datasets can be overwhelming, but a frequency distribution simplifies this by categorizing data into intervals, making it easier to understand.
2. **Data Visualization:** Frequency distribution is the foundation for creating histograms, bar charts, and frequency polygons, which are powerful tools for visualizing data.
3. **Pattern Recognition:** By observing the frequency distribution, one can identify patterns, such as the central tendency, spread, and shape of the data distribution.

Exclusive method and Inclusive method of Classifying Data

There are two methods of classifying the data according to class intervals; namely:

Exclusive Method

The exclusive method of classifying data is a thoughtful and precise approach to organizing information, particularly when dealing with continuous data that can take on any value within a specified range. Imagine you're working with data like wages, temperatures, or any other measurement where values can vary smoothly without any jumps or gaps. The exclusive method is designed to handle such data in a way that ensures clarity and avoids confusion.

In the exclusive method, class intervals are created in a way that the upper limit of one interval becomes the starting point of the next, but it doesn't belong to the first interval. For example, if you're looking at wages grouped into intervals like 40-50, 50-60, and so on, the number 50 is not included in the 40-50 interval. Instead, it belongs to the 50-60 interval.

Wages (Tk).	Number of workers
40 – 50	13
50 – 60	15
60 – 70	20
70 – 80	14
80 – 90	10

This might seem like a small detail, but it's crucial when you're dealing with continuous data. Let's say you have a group of workers and you're recording their wages. You want to know how many workers earn between 40 and 50 units of currency (let's say Tk). If you include 50 in both the 40-50 and 50-60 intervals, you'd end up counting some workers twice or creating confusion about where exactly they fall. The exclusive method prevents this by making sure each data point belongs to only one class interval.

This method is especially useful when precision is important. Imagine you're measuring something like rainfall or temperature, where values can be very specific, even down to decimal points. By excluding the upper limit of one interval and including it in the next, you create a seamless flow from one class to the next without overlap. This makes the data easier to analyze and the results more reliable.

Inclusive Method

In the inclusive method, both the lower and the upper values of a class interval are included within that interval. This means that if your class interval is 10-19, both 20 and 29 are part of this interval.

The next class interval begins with the figure immediately following the upper limit of the previous class. For example, if one class interval is 30-39, the next class interval would start at 40 and end at 49.

Number of Books	Number of Students
0 – 9	4
10 – 19	5
20 – 29	4
30 – 39	4
40 – 49	3

The inclusive method is particularly useful when dealing with discrete data, where data points are specific and separate, like counts or integers. This method ensures that each data point has an unambiguous place within the distribution, without any overlap or exclusion.

Steps to Convert Inclusive to Exclusive Class Intervals

1. Identify the Gap Between Adjacent Intervals:

The first step is to identify the gap between the upper limit of one class interval and the lower limit of the next. This gap needs to be divided equally to adjust the class boundaries.

Example: Consider the following inclusive intervals:

Class Interval (Inclusive)	Frequency
10-19	5
20-29	12
30-39	8

Here, the gap between the upper limit of 19 in the first class and the lower limit of 20 in the next class is 1 unit. We have to calculate the adjustment factor.

2. Calculate the Adjustment Factor: We have to calculate the adjustment factor.

Adjustment Factor = $\frac{\textit{Lower Limit of the 2nd Class} - \textit{Upper Limit of the First Class}}{2}$

In the above example, the gap is 1, so the adjustment factor would be:

Adjustment Factor = $\frac{20-19}{2} = 0.5$

3. Adjust the Class Limits:

We have to subtract the adjustment factor from the lower limit of each class and to add the adjustment factor to the upper limit.

Adjusting the first class interval (10-19):

Lower limit: 10−0.5=9.5
Upper limit: 19+0.5=19.5
New interval: 9.5 - 19.5

Adjusting the second-class interval (20-29):

Lower limit: 20−0.5=19.5
Upper limit: 29+0.5=29.5
New interval: 19.5 - 29.5

The new exclusive class intervals are as follows:

Class Interval (Exclusive)	Frequency
9.5 - 19.5	5
19.5 - 29.5	12
29.5 - 39.5	8

The application of the correction factor ensures continuity between the class intervals. For instance, the upper limit of one class (e.g., 29.5) matches the lower limit of the subsequent class. This ensures no gaps or overlaps between intervals, which is essential for accurate data analysis.

By converting to exclusive intervals, you avoid ambiguity in how data points on the boundaries are treated. For example, if you had a value of exactly 30, it would fall into the interval 29.5 - 39.5, rather than being ambiguously included in either 20-29 or 30-39. Exclusive intervals are commonly used in statistical analysis and data presentation, making it easier to compare datasets and maintain consistency across different analyses.

Self-Exercise

Convert the following frequency distribution with the inclusive type of class intervals into the exclusive type of class intervals.

Class Interval (C.I.)	50-59	60-69	70-79	80-89	90-99
Frequency	10	15	35	18	12

Example

Given the following ages of 35 employees:

50	52	53	55	57	58	60
62	63	65	67	68	70	71
72	73	55	56	64	66	54
51	69	58	61	62	73	74
59	57	52	60	64	66	68

Required: Form a frequency distribution taking a suitable class interval

Answer:

Determine the Range

Minimum value = 50
Maximum value = 74
Range=74−50=24

Using Sturges' formula to determine the number of classes:

$k=1+3.322\log N$

Where N is the number of observations (35):
$k=1+3.322\log 35 = 1+(3.322\times 1.544)=1+5.13\approx 6.13$
So, rounding k to the nearest whole number gives k=6
Now, to determine the class interval i,

$$i=\frac{Range}{1+3.322\ LogN}$$

$$i=\frac{maximum\ value\ -minimum\ value}{1+3.322\ LogN}$$

$$=\frac{74-5}{6}$$

$$=\frac{24}{6}=4$$

In practice, interval size is rounded up to some convenient number, such as a multiple of 5, 10, or 100.

Now, we'll create a frequency distribution table with class intervals starting at 50 and increasing by 5 (since we rounded 4 to 5 for convenience). The classes will be:

The frequency distribution table (exclusive method)

Class Interval	Tally	Frequency
50–55	𝍸 I	6
55–60	𝍸 III	8
60–65	𝍸 III	8
65–70	𝍸 II	7
70–75	𝍸 I	6
		N= 35

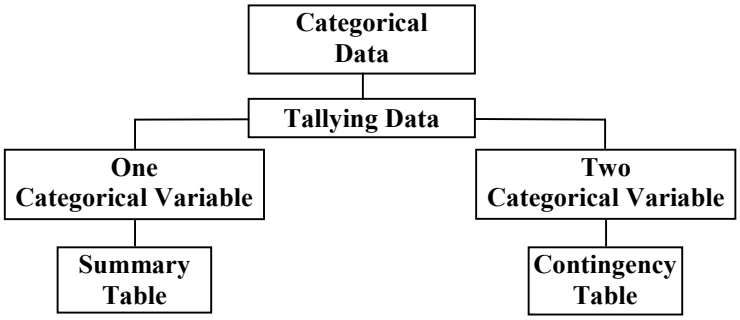
3.4 Tabulation of Data

Tabulation involves the orderly presentation of numerical facts in a tabular form to elucidate the main features of the data. It is essentially the process of condensation, simplifying and organizing raw data into an easily interpretable form.

Purpose of Tabulation:

A table is a systematic arrangement of statistical data in **columns** and **rows**. Rows are horizontal arrangements, while columns are vertical. The main objective of tabulation is to simplify data presentation and facilitate easy comparisons.

The diagram illustrates how **categorical data** can be organized using tables. It divides the data into two main approaches:



1. One Categorical Variable:

When organizing one categorical variable, we use a summary table. A summary table displays each category's frequency or count of occurrences, or percentages providing a clear overview of the data distribution for that single variable.

Modes of Transportation Used by College Students to Commute

Modes of Transportation	Percentage
Car	45%
Bicycle	15%
Public Bus	20%
Walking	12%
Other	8%

2. Two Categorical Variables:

When organizing data for two categorical variables, we use a contingency table. A contingency table (also known as a cross-tabulation table) shows the relationship between the two variables by displaying the frequency or count for combinations of categories across both variables.

Both methods start with tallying data to organize and structure the categorical information efficiently in a tabular format for analysis.

Example:

A company surveyed 300 customers to determine their satisfaction with their services. The customers were categorized into three groups based on their purchase frequency: Low, Medium, or High. They were also asked whether they were Satisfied or Not Satisfied with the service.

The data is organized into the following contingency table:

Contingency Table Showing Frequency of Customers Categorized by Purchase Frequency and Satisfaction Level

Purchase Frequency	Satisfied	Not Satisfied	Total
Low	60	20	80
Medium	90	30	120
High	80	20	100
Total	**230**	**70**	**300**

Customer Satisfaction Percentage Table

Purchase Frequency	Satisfied	Not Satisfied	Total
Low	20% (60/300)	6.67% (20/300)	26.67% (80/300)
Medium	30%	10%	40%
High	**26.67%**	6.66%	33.33%
Total	**76.67%**	**23.33%**	**100%**

For example, we can interpret the data for highly frequent purchasers of goods as indicating that they are 26.67% satisfied with their purchases.

Organizing Numerical Data - Relative & Percent Frequency Distribution

We have a dataset that shows the number of hours students spend studying in a week. The dataset is divided into the following class intervals:

Class (Hours Studied)	Frequency
5–10	4
10 - 15	7
15 - 20	5
20 - 25	3
25 - 30	1
	N= 20

Now, let's calculate the **Relative Frequency** and **Percentage** for each class interval:

1. **Relative Frequency** is calculated as:

   Relative Frequency = $\frac{Frequency\ of\ Individual\ Class}{Total\ Frequency}$

2. **Percentage** is calculated as:

   Percentage=Relative Frequency×100

Class (Hours Studied)	Frequency	Cumulative frequency	Relative Frequency	Percentage	Cumulative percentage
5 – 10	4	4	0.20	20%	20%
10 - 15	7	11	0.35	35%	55%
15 - 20	5	16	0.25	25%	80%
20 - 25	3	19	0.15	15%	95%
25 - 30	1	20	0.05	5%	100%
Total	20		1.00	100%	

The **Cumulative Frequency** column shows how many students study for up to a certain number of hours per week.

The **Relative Frequency** column represents the proportion of students in each class relative to the total number of students (20 in this case).

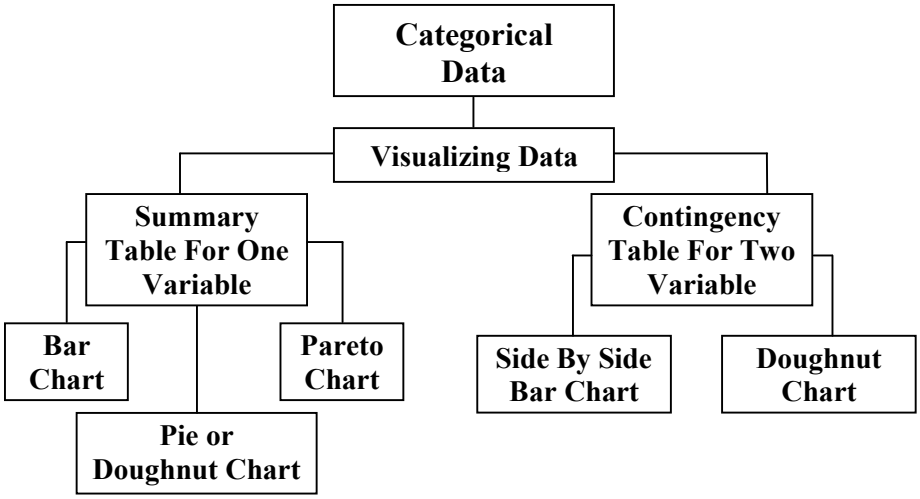
The **Percentage** column converts the relative frequency into a percentage for easier interpretation.

The **Cumulative Percentage** column gives the percentage of people who fall within that cumulative frequency, making it easier to visualize how the data accumulates over different intervals.

This table shows how study hours are distributed among students, with most students studying between 10 and 15 hours per week.

3.5 Charting Data

Visualizing Categorical Data Through Graphical Displays



Bar Chart

A bar chart is a type of graphical display used to represent data with rectangular bars where the length of each bar is proportional to the value it represents. It's particularly useful for comparing quantities across different categories.

Example

Let's say we want to visualize the monthly expenses of a household in Dhaka for different categories in Bangladeshi Taka (BDT). Here are some example data points:

- Rent: 15,000 BDT
- Groceries: 7,000 BDT
- Utilities: 3,000 BDT
- Transportation: 2,500 BDT
- Entertainment: 1,500 BDT

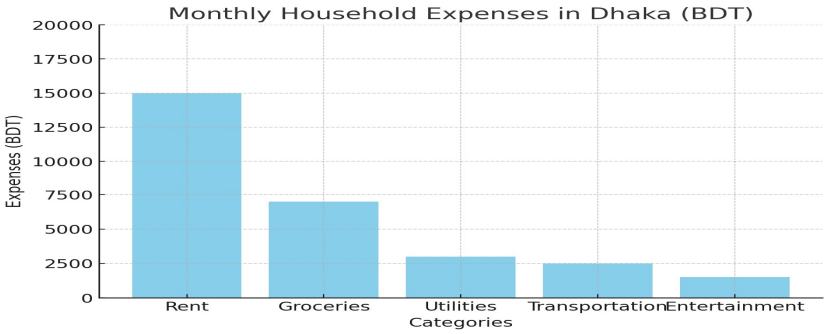


Figure: Bar Chart

Here's the bar chart depicting monthly household expenses in Dhaka, represented in Bangladeshi Taka (BDT). Each bar reflects the amount spent on different categories such as rent, groceries, utilities, transportation, and entertainment. This visualization helps in easily comparing the expenditures across these categories, showing that rent is the largest expense, followed by groceries and utilities.

The Pareto Chart

The Pareto chart is a specialized type of bar chart where values are represented in descending order of relative frequency from left to right. Additionally, it includes a line graph that shows the cumulative total percentage. Pareto charts are particularly useful for identifying the most significant factors in a set of data and are often used in quality control to prioritize issues that need attention based on their significance.

Example

The table shows the customer complaints about the services of a company. The graph shows the Pareto chart of the data.

Customer Complaint types	Frequency	Cumulative frequency	Cumulative percentage
Long wait times	450	450	64%
Poor food quality	100	550	79%
Unfriendly staff	50	600	86%
Incorrect orders	45	645	92%
Uncomfortable seating	30	675	96%
High prices	25	700	100%

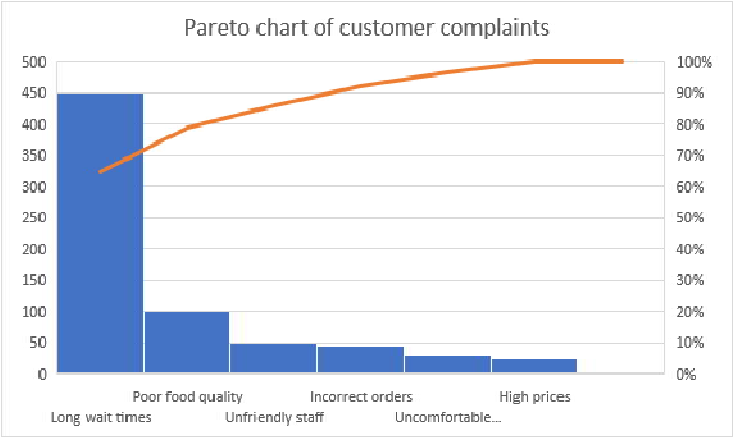


Figure: Pareto chart

Pie chart

A pie chart is a circular chart divided into slices to illustrate numerical proportions. Each slice represents a category, and the size of the slice is proportional to its share of the total.

Example

Suppose we want to visualize the distribution of expenses for a project:

Expenses Area	Percentage of expenses of total Budget
Research	40%
Development	30%
Marketing:	20%
Operations	10%

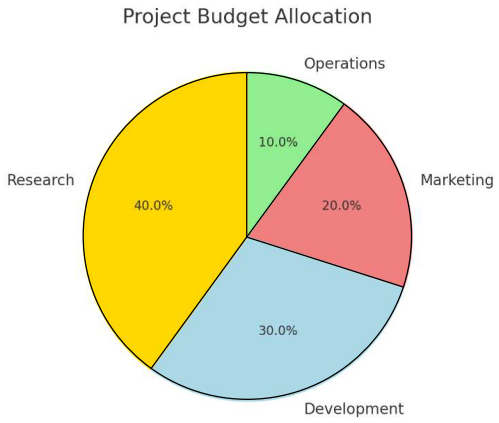


Figure: Pie Chart

Here is the pie chart illustrating the project budget allocation across four categories: Research, Development, Marketing, and Operations. Each slice represents the proportion of the total budget allocated to each category, with the percentages clearly labelled. This type of chart is useful for visualizing the distribution of resources or other categorical data.

Side-by-side Bar Chart

A side-by-side bar chart (also known as a grouped bar chart) is used to compare two or more sets of data across different categories. It displays multiple bars for each category, where each bar represents a different data set, allowing for easy comparison between the data sets within each category.

Example

The following table shows sales data for two different stores (Store A and Store B) across three months:

Months	Store A	Store B
January	Tk. 20,000	Tk. 18,000
February	25,000	30,000
March	22,000	20,000

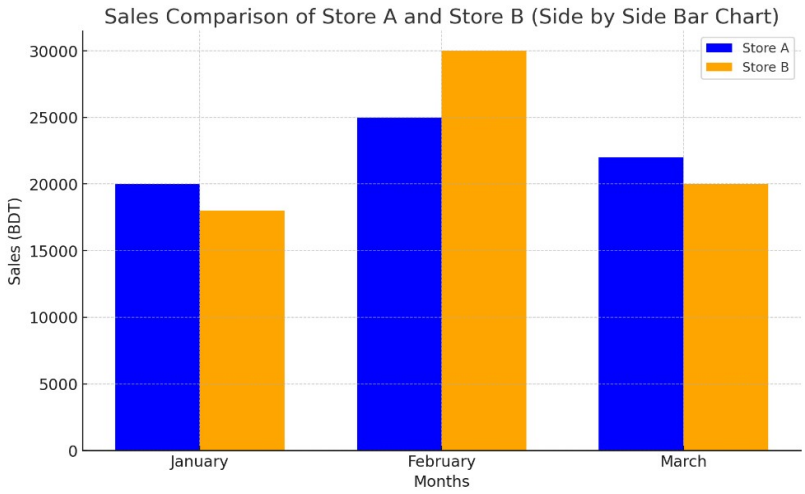


Figure: Side-by-side bar chart

Here is the side-by-side bar chart comparing sales between Store A and Store B over three months (January, February, and March). Each pair of bars represents the sales for each month, making it easy to compare the sales performance of the two stores for each month.

Doughnut Chart

A doughnut chart is similar to a pie chart but with a hole in the center. It is often used to represent proportions, and its central hole can be used to add a second layer of data.

The table shows the market share of four tech companies:

Company	Market Share
A	40%
B	30%
C	20%
D	10%

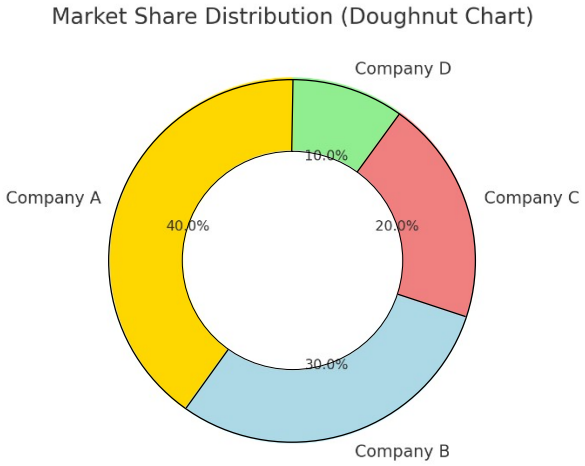


Figure: Doughnut Chart

Here is the doughnut chart representing the market share distribution of four tech companies. Each slice of the chart corresponds to the proportion of market share held by the respective

company, with **Company A** holding the largest share at 40%, followed by **Company B** at 30%, **Company C** at 20%, and **Company D** at 10%.The central hole distinguishes this from a pie chart and helps visually declutter the chart.

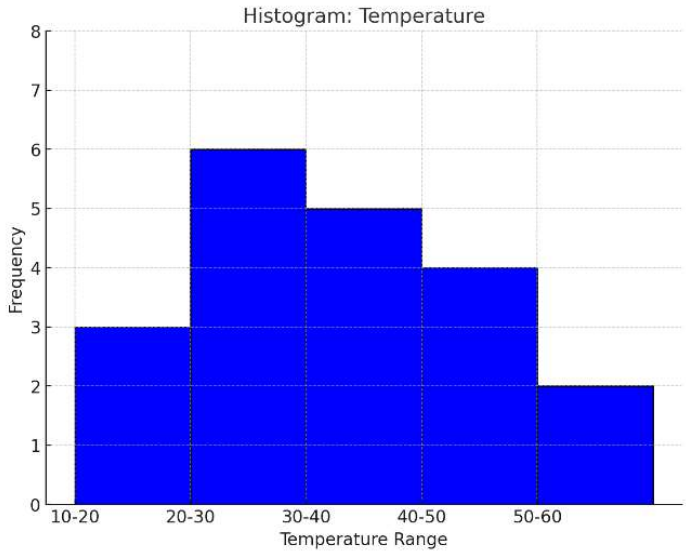
Visualizing numerical data using graphical displays

It can be done through:

1. Histogram
2. Cumulative Polygon or Ogive

1. Histogram

A histogram is a vertical bar chart used to visualize the frequency distribution of numerical data. Unlike bar charts, histograms do not have gaps between adjacent bars. The x-axis represents class boundaries (or midpoints), which group the data into bins. The y-axis can represent frequency, relative frequency, or percentage of occurrences within each bin. The height of each bar shows how many data points fall within that bin (the frequency or percentage).



2. Cumulative Polygon or Ogive

A Cumulative Percentage Polygon, also known as an Ogive, is a graphical representation used in statistics to display a dataset's cumulative frequencies or relative frequencies.

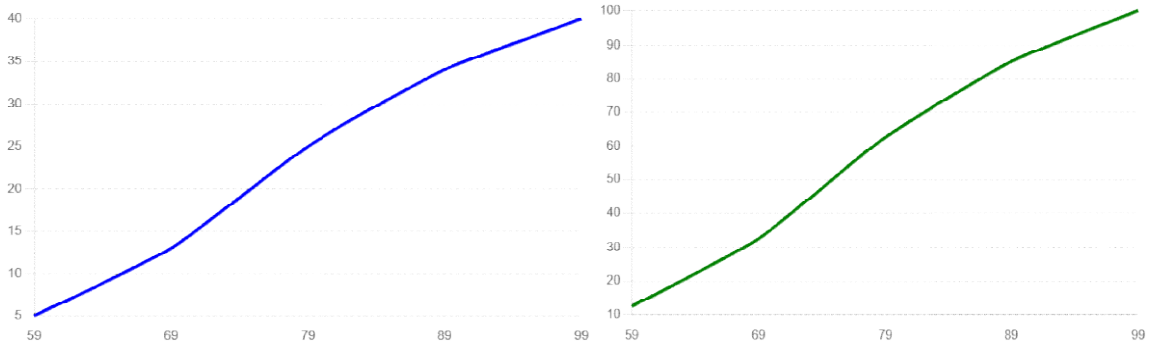


Figure: Cumulative Frequency Polygon (Ogive) and the Cumulative Percentage Polygon (Ogive)

Review Questions

1. Explain the purpose of data presentation and how it aids in data analysis.
2. Differentiate between qualitative and quantitative data classification with examples.
3. What is frequency distribution, and why is it useful in data analysis?
4. Describe the exclusive and inclusive methods of data classification. Give examples of each.
5. What are the key differences between a bar chart and a histogram? In what scenarios would each be used?
6. Explain the Pareto chart and discuss its application in identifying significant data categories.
7. What is the purpose of tabulation, and how does it help in organizing numerical data?
8. Describe a cumulative polygon (ogive) and explain how it helps in understanding data distribution.
9. When should a pie chart be used over a bar chart? Provide an example scenario.
10. What is an array in data classification, and how does it simplify data analysis?
11. The following data represents the number of hours 30 students spent on homework in a week.
    3, 4, 5, 5, 6, 7, 8, 9, 10, 10, 11, 12, 12, 13, 14, 14, 15, 16, 17, 18, 18, 19, 20, 21, 22, 23, 24, 25, 25, 26
    Construct a frequency distribution table for the given data.
12. A retail store's sales department studies a product's daily sales over the past month. A sample of 30 days of sales data for this product is recorded as follows:
    \$135, \$148, \$112, \$125, \$157, \$132, \$139, \$142, \$128, \$115,
    \$160, \$168, \$154, \$125, \$130, \$138, \$122, \$145, \$136, \$158,
    \$140, \$127, \$123, \$150, \$156, \$133, \$131, \$149, \$147, \$134.

    Requirement: a. Organize the data into a frequency distribution with the exclusive method. Find the suitable classes and class intervals.
    b. Draw a histogram from this frequency distribution.
13. The data below gives the yearly profits (in thousands of Taka) of two companies, A and B.

Year	*Profits in ('000 taka)*	
	Company A	*Company B*
2014-15	120	90
2015-16	135	95
2016-17	140	115
2017-18	160	120
2018-19	170	130

    Represent the data with a suitable diagram.
14. Given the data below, create a cumulative frequency polygon:

Score Range	Frequency
10-19	3
20-29	7
30-39	10
40-49	5

15. Given the frequency table below, calculate the cumulative frequency:

Age Group	Frequency
10-19	5
20-29	8
30-39	12
40-49	10

References

Abdullah, A. (2019). Data Classification. *CCSP (ISC)2 Certified Cloud Security Professional Official Study Guide, 2nd Edition*. https://doi.org/10.1002/9781119603351.ch3.

Aggarwal, C. C., & Aggarwal, C. C. (2015). *Data classification* (pp. 285-344). Springer International Publishing.

CENTRAL TENDENCY

4

Unit Highlights

- Introduction
- Mean
- Median
- Mode
- Relationship among mean, median, and mode
- Measures of Relative Position

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 4: Central Tendency

Learning Objectives

By the end of this Unit, you should be able to:

1. Define and understand the concept of central tendency and its role in summarizing data.
2. Compute the arithmetic mean, median, and mode for both ungrouped and grouped data using various methods.
3. Explain the characteristics, uses, and limitations of the mean, median, and mode as measures of central tendency.
4. Explore the relationship among mean, median, and mode and apply this understanding to analyze data distributions.
5. Calculate measures of relative position, including quartiles, deciles, and percentiles, to interpret data effectively.

4.1 Introduction

Central tendency refers to a single value that describes the center of a data set, representing the typical or most frequent value around other values in the data cluster (Sial & Abid, 2023). It is a fundamental concept in statistics that helps summarize large datasets by identifying a central position within the data (Anilkumar, 2013). Measures of central tendency, which include the mean, median, and mode, provide insights into the middle point of the data distribution (Chakrabarty, 2021). These measures are often referred to as measures of central location because they point out where the center of the data lies. The mean or average, is the most commonly known measure and is calculated by adding all the values together and dividing by the number of observations. Central tendency is essential in statistics as it offers a simplified way to understand complex datasets by focusing on one representative value that encapsulates the overall trend or center of the data. Here's a detailed look at the concept of central tendency:

Importance of Central Tendency

Central tendency simplifies the complexity of large data sets by providing a single representative number. It allows comparing different data sets with a single summary statistic. In business, science, and data analysis, decisions are often based on the central values of observed data.

Measures of Central Tendency

The three primary measures of central tendency are the mean, median, and mode, each of which provides a different perspective on the data:

1. Mean
2. Median
3. Mode

4.2 Mean

The mean, commonly referred to as the average, is one of the most widely used measures of central tendency. It applies to both discrete and continuous data, though it is more frequently used with continuous data. The mean is calculated by adding all the values in a dataset and dividing the sum by the total number of observations. This gives a single representative value

that indicates the central point of the data. Mathematically, the formula for the mean can be expressed as the sum of all observations divided by the number of observations.

4.2.1 Arithmetic Mean

The arithmetic mean is the most commonly used type of mean, and it is often simply referred to as "the mean" or "the average." It is calculated by adding all the values in a dataset and then dividing by the total number of values. The arithmetic mean is widely used in various fields like economics, finance, and everyday decision-making (Meyer et al., 1995).

Arithmetic Mean for Ungrouped Data

The arithmetic mean of ungrouped data is calculated by taking the sum of all individual data points and dividing it by the total number of data points. Ungrouped data refers to raw data that has not been organized into groups or categories. The formula for the arithmetic mean of ungrouped Data:

Arithmetic mean, $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

Where:

x_i represents each data point

n is the total number of data points

$\sum$ symbolizes the sum of all values in the dataset

For example, a dataset showing the scores of 10 students on a test:

85, 90, 78, 88, 92, 75, 89, 95, 81, 87

The sum of all scores: $\sum x_i$= 85+90+78+88+92+75+89+95+81+87=860

Total number of students: n=10

Hence, arithmetic mean: $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{860}{10} = 86$

Arithmetic Mean for Grouped Data

The arithmetic mean for grouped data is slightly more complex than for ungrouped data because the data is organized into groups or classes that we have learned in the frequency distribution. To calculate the mean for grouped data, we use the midpoint (also called class mark) of each group and the frequencies of the groups.

The formula for Arithmetic Mean of Grouped Data:

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum f_i} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \ldots\ldots + f_n}$$

Where:

- f_i = frequency of each class (number of observations in the group).
- x_i = midpoint (or class mark) of each class, which is calculated as $\frac{Lower\ limit + Uppe\ \ limit}{2}$
- $\sum f_i$ = sum of frequencies, as we denote it by n also

Here's an example dataset of student scores grouped into intervals:

Student scores	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100
No. of persons	3	7	8	5	2

Solution:

Find the midpoints (x_i) for each class and $f_i x_i$:

Class Interval	Midpoint (x_i)	Frequency f_i	$f_i x_i$
50 – 60	55	3	165
60 – 70	65	7	455
70 – 80	75	8	600
80 – 90	85	5	425
90 – 100	95	2	190
		$n \text{ or } \sum f_i = 25$	$\sum f_i x_i = 1835$

Hence, the arithmetic mean, $\overline{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum f_i} = \frac{1835}{25} = 73.4$

Short-cut method for Arithmetic Mean of Grouped Data using assumed mean:

The shortcut method (or assumed mean method) is a quicker way to calculate the arithmetic mean for grouped data, especially when the numbers are large. This method reduces the complexity of calculations by using an "assumed mean" to simplify the calculations of deviations.

The formula for the mean of the short-cut method:

$$\overline{X} = A + \frac{\sum f_i d_i}{n \text{ or } \sum f_i} \times i$$

Where:

- A = Assumed mean (a value taken from the midpoints of one of the classes).
- f_i = Frequency of each class.
- $i =$ the size of the class interval
- $d_i = \frac{x_i - A}{i}$ Deviation of the class midpoint (x_i) from the assumed mean divided by class interval.
- $\sum f_i d_i$ = Sum of the product of frequencies and deviations.
- n or $\sum f_i$ = Total frequency.

Let's assume the assumed mean A= 75,

Class Interval	Midpoint (x_i)	$d_i = \frac{x_i - A}{i} = \frac{x_i - 75}{10}$	Frequency f_i	$f_i d_i$
50 – 60	55	-2	3	-6
60 – 70	65	-1	7	-7
70 – 80	75	0	8	0
80 – 90	85	1	5	5
90 – 100	95	2	2	4
			$n \text{ or } \sum f_i = 25$	$\sum f_i d_i = -4$

The arithmetic mean would be,

$$\overline{X} = A + \frac{\sum f_i d_i}{n \text{ or } \sum f_i} \times i$$

$= 75 + \frac{-4}{25} \times 10$

$= 75 - 1.6$

$= 73.4$

The Formula for Population Mean

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

Where:
μ = Population Mean (the Greek letter "*mu*")
x_i = Each data point in the population.
N = Total number of data points in the population.
$\sum$ = Sum of all the data points.

4.2.2 Characteristics and Uses of Mean

1. Sensitive to Outliers: The mean takes into account every value in the dataset, which means that extreme values (outliers) can significantly affect the mean. A few very high or very low values can skew the mean away from the central tendency of most of the data points.

Suppose we have the following test scores for 5 students:

85, 90, 88, 92, 87
Mean = $\frac{85+90+88+92+}{5}$ = 88.4

So, the mean score without any outliers is 88.4.

Now, let's say there was an error in recording one student's score, and it was mistakenly recorded as 40 instead of a higher score, creating an outlier: 85, 90, 88, 92, 40

Calculate the mean with the outlier:

Mean = $\frac{85+90+88+92}{5}$ = 79

Without the outlier, the mean was 88.4, where the mean dropped significantly to 79.

This example shows that the mean is highly sensitive to outliers, as just one extreme value (40) drastically lowered the overall mean. If you used the mean to summarize this data, it might give a misleading impression that the scores are much lower overall.

2. Applicable to Interval and Ratio Data: The mean can be calculated for data measured on an interval or ratio scale, where the distance between values is meaningful (e.g., height, weight, temperature). It is not typically used for nominal or ordinal data, where the data represents categories or ranks.

3. Unique Value: The mean is a unique value for a given dataset, unlike the mode, which might have multiple values in a multimodal distribution.

4. Affected by All Data Points: The mean is computed using all the data points in a dataset. This characteristic ensures that the mean reflects every value, but it also means it might not represent the dataset well if the data is skewed.

Example of error correction of arithmetic mean

Illustration: The mean of 100 observations was originally calculated as 40. Later, it was discovered that two observations were wrongly read as 28 and 46, instead of the correct values of 82 and 146. Calculate the correct mean.

Original Sum of the Observations**:**

$$\sum X = N \times \overline{X} = 100 \times 40 = 4{,}000$$

Subtract the incorrect values (28 and 46):4,000−(28+46)=4,000−74=3,926
Add the correct values (82 and 146): 3,927+(82+146)=4,154
Divide the corrected total sum by the number of observations (100):
Correct Mean = $\frac{4,154}{100}$ = 41.54

4.2.3 Weighted Arithmetic Mean

The weighted arithmetic mean is an average where each data point is assigned a specific weight. It is used when certain values in a dataset have more importance (weight) than others. The weights reflect the relative significance of each data point in contributing to the overall average.

The formula for Weighted Arithmetic Mean:

$$\overline{X} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i}$$

Where:

- w_i = the weight of each observation.
- x_i = each data point.
- $\sum w_i$ = the sum of the weights.

Illustration:

A teacher wants to calculate the final grade for a student based on their performance in several categories: homework, quizzes, two projects, and a final exam. The teacher decides to weight the grades as Final Exam: 50%, Projects: 25% (each project counts as 12.5%), Quizzes: 15%, Homework: 10%

The student's grades in each category are as Final Exam: 78, Project 1: 85, Project 2: 90, Quizzes: 72, Homework: 88.

Calculate the student's final weighted average for the course.

Solution:

Weights:

Final Exam**:** 50% or .50
Project 1**:** 12.5% or .125
Project 2**:** 12.5% or .125
Quizzes**:** 15% or .15
Homework**:** 10% or .10

Multiply each grade by its respective weight:

Final Exam: 78×0.50=39
Project 1: 85×0.125=10.625
Project 2: 90×0.125=11.25
Quizzes: 72×0.15=10.8
Homework: 88×0.10=8.8
Weighted average, $\overline{X} = \frac{39+\ \ .625+1\ \ .25+10.8+8.8}{1}$ = 80.475

4.2.4 Geometric Mean

The geometric mean is another type of mean, which is especially useful for datasets involving ratios, percentages, or exponentially growing quantities. It is calculated by

multiplying all the values together and then taking the nth root, where n is the total number of values in the dataset. The geometric mean is often used in fields like finance, biology, and environmental science.

$$GM = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Where,x_i = each data point

n = the number of terms

$\prod_{i=1}^{n} x_i$= products of all growth factor

Illustration: Consider the growth of Tk.2,000 deposited in a savings account over 4 years. The interest rates for the 4 years are as follows:

- Year 1: 6%
- Year 2: 9%
- Year 3: 11%
- Year 4: 15%

The growth factor for each year is calculated as $1+\frac{interest\ rate}{100}$

Requirements:

1. Calculate the growth factor for each year.
2. Using the growth factors, calculate the geometric mean.
3. Using the geometric mean, estimate the overall growth factor over the 4 years.
4. What would be the amount in the account at the end of 4 years?

Solution:

- Initial deposit: Tk.2,000
- Interest rates:
  - Year 1: 6%
  - Year 2: 9%
  - Year 3: 11%
  - Year 4: 15%

We have to calculate the growth factor for each year

1. The growth factor is given by $1+\frac{interest\ rate}{100}$

For each year:

- Year 1: $1+\frac{6}{100}$ =1.06
- Year 2: $1+\frac{9}{100}$ =1.09
- Year 3: $1+\frac{11}{100}$ =1.11
- Year 4: $1+\frac{15}{100}$ =1.15

2. The formula for the geometric mean is:

Geometric Mean $= (1.06 \times 1.09 \times 1.11 \times 1.15)^{\frac{1}{4}}$

= 1.1020.

The average annual interest rate is 10.20%

3. Overall Growth Factor $= (1.1020)^4$
$= 1.475$

4. The final amount in the account after 4 years is:

Final Amount=Initial Deposit × Overall Growth Factor
Final Amount=2,000×1.475=Tk. 2,950.00
Thus, the amount in the account at the end of 4 years will be approximately Tk.2,950.00

4.3 Median

The median is a measure of central tendency that identifies the middle value in a dataset when the values are arranged in ascending or descending order. It is especially useful when a dataset contains outliers, as it is not affected by extreme values like the mean.

Steps to Calculate the Median:

Arrange the data in ascending order (from smallest to largest).
If the number of data points (n) is odd, the median is the middle value.

Median = Size of $\frac{N+1}{2}th$ observation

If the number of data points (n) is even**,** the median is the average of the two middle values.

Example: Odd Number of Data Points

Let's take the following dataset of 7 student scores:

65, 70, 75, 80, 85, 90, 95

Since there are 7 data points (odd number), the median is the middle value $\frac{7+1}{2}th$= 4th observation

Median = 80 (the 4th observation).

Even Number of Data Points
Now, consider this dataset of 6 student scores:

60, 65, 70, 75, 80, 85

Since there are 6 data points (even number), the median is the middle value $\frac{6+1}{2}th$= 3.5th observation; thus, the median would be the average of the two middle values.
Middle values between 70 and 75.
$\frac{70+}{2}$ = 72.5
So, the median is 72.5.

Key Points:

- The median is useful when dealing with skewed data or data with outliers, as it provides a better representation of the central value in such cases.

- Unlike the mean, the median is not affected by extremely high or low values.

Median for Group Data

The median for grouped data is calculated using the frequency distribution table, where the data is grouped into class intervals. To find the median for grouped data, we use the following formula:

Median $(M_d) = L + \frac{\frac{N}{2} - p.c.f}{f} \times i$

Where:

L = Lower boundary of the median class.
N = Total number of frequencies (sum of all frequencies).
$p.c.f$ = Preceding Cumulative Frequency to the median class.
f = Frequency of the median class.
i = Class width (size of the median class interval).

To calculate the median:

First, divide the total number of frequencies by 2. Then, find the median class where the cumulative frequency just exceeds $\frac{N}{2}$.
Apply the values to the median formula to calculate the median.

Illustration:

From the following data on the weights (in kg) of 150 persons, determine the median weight using the method of grouping:

Weight (in kg)	40-50	50-60	60-70	70-80	80-90	90-100	100-110	110-120	120-130
No. of Persons	3	7	15	30	35	25	18	10	7

Solution:

Weight	Midpoint (X)	Frequency (f)	Cumulative frequency ($c.f$)
40-50	45	3	3
50-60	55	7	10
60-70	65	15	25
70-80	75	30	55
80-90	85	35	90
90-100	95	25	115
100-110	105	18	133
110-120	115	10	143
120-130	125	7	150

To find the median weight, we have to determine the median class.
Total number of persons (N) = 150
Median lies on $\frac{150}{2}$ th observation, 75th observation. So, the median class is 80-90 because 75 lies between the cumulative frequency of 55 and 90.

Median = $L + \frac{\frac{N}{2} - p.c.f}{f} \times i$

L= 80
pcf = 55 (cumulative frequency before the median class).
f = 35 (frequency of the median class).
i = 10 (class width).

Median (M_d) = $80 + \frac{\frac{150}{2} - 55}{35} \times 10$
= 80+5.71
= 85.71 kg

The median weight for this dataset is approximately 85.71 kg. This method provides an estimated median for grouped data using cumulative frequencies and class intervals.

4.4 Mode

The mode is a measure of central tendency that represents the value or values that occur most frequently in a dataset. It is the number that appears with the highest frequency. Unlike the mean and median, the mode can be used with both numerical and categorical data.

Key Characteristics of the Mode:

1. **Most Frequent Value**: The mode is the value that appears most often in a dataset.
2. **Multiple Modes (Bimodal, Multimodal)**: A dataset can have more than one mode:
   Unimodal: Only one mode (one value that appears most frequently).
   Bimodal: Two values appear with the highest frequency.
   Multimodal: More than two values appear with the highest frequency.
3. **No Mode**: If no value repeats, the dataset does not have a mode.
4. **Not Affected by Extreme Values**: The mode is not influenced by outliers or extreme values.

Example 1: Numerical Data (Unimodal)

Consider the following dataset of test scores:
Scores: 55, 70, 65, 70, 85, 90
The value 70 appears twice, more frequently than any other value.
Mode = 70

Example 2: Numerical Data (Bimodal)

Consider the following dataset of ages:
Ages: 22, 25, 25, 30, 30, 35, 40
Both 25 and 30 appear twice, which is the highest frequency.
Mode = 25 and 30 (Bimodal)

Example 3: Categorical Data

Consider the following dataset of favorite colors:
Colors: Red, Blue, Blue, Green, Red, Blue
The color Blue appears three times, more frequently than any other color.
Mode = Blue

The Usefulness of the Mode

The mode is useful in determining the most common or popular item in a dataset, such as finding the most frequent product purchased, the most common exam score, or the most popular color preference. It can be applied to **categorical data** also (e.g., favorite fruit, preferred color), whereas the mean and median do not apply to such data.

Limitations

The mode may not provide useful information if the data is uniformly distributed (no value repeats). In cases of multimodal data, summarizing with a single mode may not accurately describe the dataset.

Mode for Group Data

To find the mode, we have to identify the modal class. The modal class is the class with the highest frequency.

The formula for the mode is:

Mode(M_o)=$L+\frac{f_1-f_0}{(f_1-f_0)+(f_1-f_2)}\times i$

Where:
L= lower boundary of the modal class.
f_1= frequency of the modal class.
f_0= frequency of the class before the modal class.
f_2=frequency of the class after the modal class.
i=class width

Illustration:

Consider the data above example of the median.

Weight	Midpoint(X)	Frequency (f)
40-50	45	3
50-60	55	7
60-70	65	15
70-80	75	30
80-90	85	35
90-100	95	25
100-110	105	18
110-120	115	10
120-130	125	7

Find the modal weight of the frequency distribution.

Answer:

To find out the modal weight of this frequency table, first identify the Modal Class. From the table, the class 80-90 kg has the highest frequency of 35, so this is the modal class.

The formula for the mode is:

Mode (M_o) = $L + \frac{f_1-f_0}{(f_1-f_0)+(f_1-f_2)}\times i$

Where:

L=80 (lower boundary of the modal class).
f_1=35 (frequency of the modal class).
f_0=30 (frequency of the class before the modal class, 70-80 kg).
f_2=25 (frequency of the class after the modal class, 90-100 kg).
i=10 (class width).

Putting the value in the formula,

Mode (M_o) = $80 + \frac{35-30}{(35-30)+(35-25)}\times 10$

= 80+3.33

The modal weight for this dataset is approximately 83.33 kg. The mode represents the most frequent value in a dataset, and for grouped data, it is calculated using the frequencies of the modal class and the neighboring classes.

If there are more than one modal class, then the formula to calculate the Mode,
Mode(M_o)=3×Median(M_d)−2×Mean(M)

4.5 Relationship among mean, median, and mode

The relationship among mean, median, and mode is fundamental to understanding the shape and distribution of a dataset. These three measures of central tendency each provide a different perspective on where the "center" of the data lies, and their relationship helps describe the nature of the data's distribution.

In a symmetrical distribution, where the data is evenly spread on both sides of the center, the mean, median, and mode all coincide at the same point. This is often seen in a normal distribution or bell curve, where the values cluster around the middle, and there are fewer occurrences of extremely low and high values. In such a distribution, the mean is equal to the median and the mode. This relationship is common in many natural phenomena like heights, test scores, and IQ distributions, where most data points tend to cluster around the average value. For example, in a group of students' test scores, if most scores are centered around a typical value with fewer outliers, the mean, median, and mode will be the same, reflecting a balanced, symmetrical pattern.

However, not all datasets are symmetrical. In many cases, we encounter skewed distributions, where extreme values pull the data to one side.

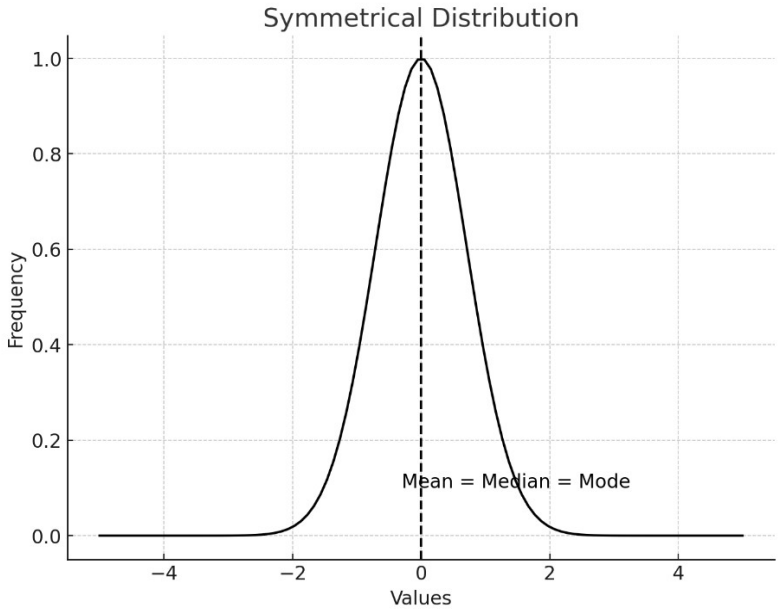
In a positively skewed distribution (or right-skewed distribution), the tail extends more to the right, indicating that there are a few unusually high values pulling the average upward. Here, the mean will be greater than the median, which in turn will be greater than the mode. This is common in income data, where most people earn a moderate amount, but a small number of very high incomes raise the mean. For example, if we look at income levels in a country, the mode (most common income) and the median (middle income) may be much lower than the mean, which is skewed upward by a few high-income individuals. This creates a right-skewed pattern.

In contrast, a negatively skewed distribution (or left-skewed distribution) has a long tail on the left side, meaning that a few low values drag the mean down. In this case, the mode is greater than the median, and the mean is the smallest of the three. This type of distribution can occur when there is a limit on how low values can go but no limit on how high they can go, such as in certain types of performance scores or reaction times where most people perform well but a few outliers struggle. For instance, in an easy exam, most students may score high marks, but a few lower scores could pull the mean down, creating a left-skewed distribution. Here, the mode represents the common high score, the median is slightly lower, and the mean is the lowest, pulled down by a few outliers.

To summarize, the relationship between the mean, median, and mode tells us a great deal about the distribution of the data:

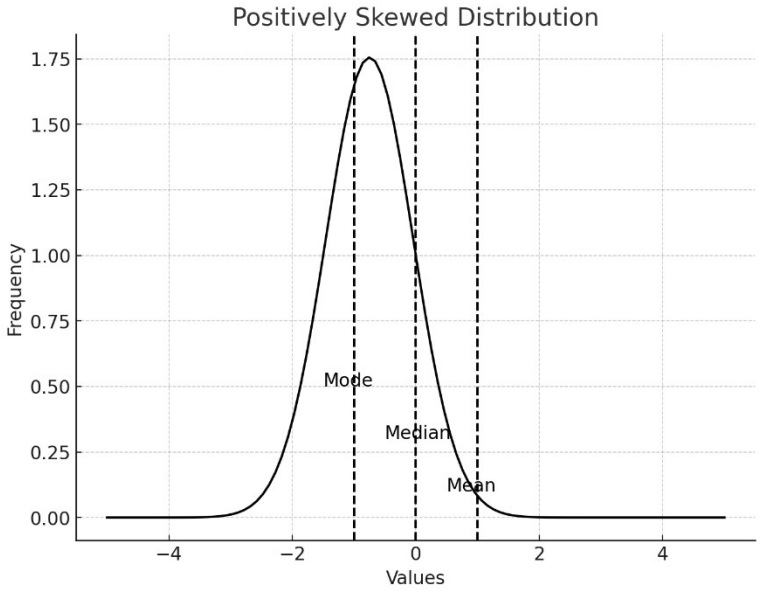
Symmetrical Distribution:

In a symmetrical distribution, the mean, median, and mode all coincide at the center of the distribution. This is typical of normal distributions where data is evenly distributed on both sides.



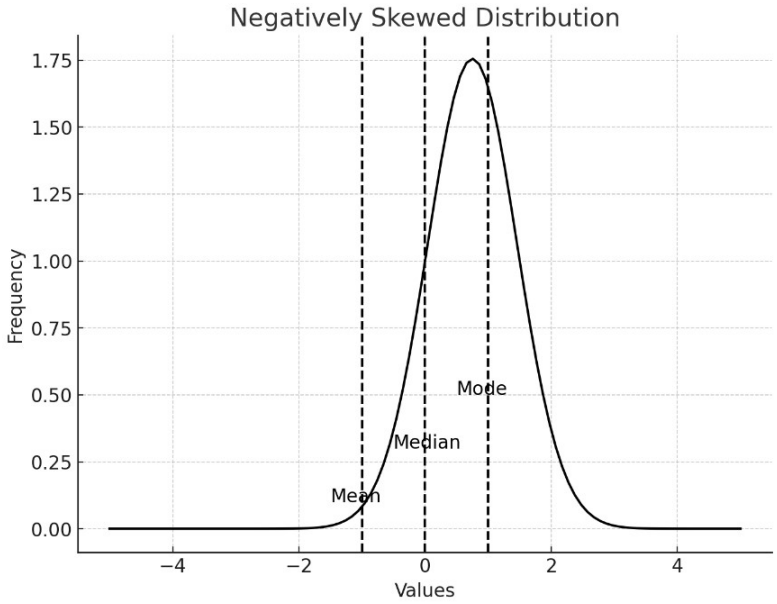
Positively Skewed Distribution:

In a positively skewed distribution, the mean is greater than the median, and the median is greater than the mode. The tail extends to the right, pulling the mean further to the right of the mode and median.



Negatively Skewed Distribution:

In a negatively skewed distribution, the mode is greater than the median, and the median is greater than the mean. The tail of the distribution extends to the left, indicating that there are a few lower values pulling the mean downward.



4.6 Measures of Relative Position

4.6.1 Quartile

These divide a dataset into 4 equal parts. The three quartile values (Q_1, Q_2, and Q_3) correspond to the 25th percentile, the median (50th percentile), and the 75th percentile, respectively.

4.6.2 Decile

These divide a dataset into 10 equal parts. Each decile represents 10% of the distribution. For example, the first decile represents the lowest 10% of the data, while the ninth decile represents the top 10%.

4.6.3 Percentiles

These divide a dataset into 100 equal parts. A specific percentile (e.g., the 90th percentile) indicates the value below which a given percentage of observations fall. For example, if a score is in the 90th percentile, it means the score is higher than 90% of the other scores.

Let's solve these by using the data

Example: From the following data on the weights (in kg) of 150 persons,

1. Find the quartiles Q_1, Q_2, Q_3
2. 3^{rd}, 7th deciles, and
3. 60^{th}, 80^{th} percentiles.

Weight (in kg)	40-50	50-60	60-70	70-80	80-90	90-100	100-110	110-120	120-130
No. of Persons	3	7	15	30	35	25	18	10	7

Solution: Let's do the frequency table.

Weight	Midpoint (X)	Frequency (f)	Cumulative frequency ($c.f$)
40-50	45	3	3
50-60	55	7	10
60-70	65	15	25
70-80	75	30	55
80-90	85	35	90
90-100	95	25	115
100-110	105	18	133
110-120	115	10	143
120-130	125	7	150

The quartiles Q_1, Q_2, and Q_3.

The first quartile lies on $\frac{N}{4} = \frac{150}{4} = 37.5^{th}$ observation. The first quartile lies in the class 70-80 because 37.5 lies between the cumulative frequency of 25 and 55.

$$Q_1 = L + \frac{\frac{N}{4} - p.c.f}{f} \times i$$
$$= 70 + \frac{\frac{150}{4} - 25}{30} \times 10$$
$$= 70 + 4.17 = 74.16$$

The second quartile lies on $\frac{2N}{4} = \frac{300}{4} = 75^{th}$ observation. The second quartile lies in the class 80-90 because 75 lies between the cumulative frequency of 55 and 90.

$$Q_2 = L + \frac{\frac{2N}{4} - p.c.f}{f} \times i$$
$$= 80 + \frac{\frac{2*150}{4} - 55}{35} \times 10$$
$$= 80 + 5.71 = 85.71$$

The third quartile lies on $\frac{3N}{4} = \frac{450}{4} = 112.5^{th}$ observation. The third quartile lies in the class 90-100 because 112.5 lies between the cumulative frequency of 90 and 115.

$$Q_3 = L + \frac{\frac{3N}{4} - p.c.f}{f} \times i$$
$$= 90 + \frac{\frac{3*150}{4} - 90}{25} \times 10$$
$$= 90 + 9 = 99$$

3rd, 7th deciles

The 3rd decile lies on $\frac{3N}{10} = \frac{450}{10} = 45^{th}$ observation. The 3rd decile lies in the class 70-80 because 45 lies between the cumulative frequency of 25 and 55

$$D_3 = L + \frac{\frac{3N}{10} - p.c.f}{f} \times i$$
$$= 70 + \frac{\frac{3*150}{10} - 25}{30} \times 10$$
$$= 70 + 6.67$$
$$= 76.67$$

The 7th decile lies on $\frac{7N}{10} = \frac{7*150}{10} = 105^{th}$ observation. The 7th decile lies in the class 90-100 because 105 lies between the cumulative frequency of 90 and 115

$$D_7 = L + \frac{\frac{7N}{10} - p.c.f}{f} \times i$$
$$= 90 + \frac{\frac{7*150}{10} - 90}{25} \times 10$$
$$= 90 + 6$$
$$= 96$$

The 60th percentile lies on $\frac{60N}{100} = \frac{60*150}{100} = 90^{th}$ observation. The 60th percentile lies in the class 80-90 because 90 lies between 55 and 90

$$P_{60} = L + \frac{\frac{60N}{100} - p.c.f}{f} \times i$$
$$= 80 + \frac{\frac{60*150}{100} - 55}{35} \times 10$$
$$= 80 + 10$$
$$= 91$$

The 80th percentile lies on $\frac{80N}{100} = \frac{80*150}{100} = 120^{th}$ observation. The 80th percentile lies in the class 100-110 because 120 lies between the cumulative frequency of 115 and 133

$$P_{80} = L + \frac{\frac{80N}{100} - p.c.f}{f} \times i$$
$$= 100 + \frac{\frac{80*150}{100} - 115}{18} \times 10$$
$$= 100 + 2.78$$
$$= 102.78$$

Review Questions

1. Define central tendency. Why is it important in statistics?
2. Explain the mean, median, and mode. How are they different from each other?
3. In a data set where most values are the same, what can be said about the mode?
4. Under what conditions is the median more suitable than other measures of central tendency?
5. A student scored the following marks in six subjects: 72, 85, 90, 75, 88, 79. Calculate the mean score.
6. In a survey, the number of hours students' study in a week are as follows: 8, 10, 7, 8, 10, 6, 7, 9. Find the mode.
7. Explain why the mean can be misleading in a data set with extreme values.
8. Calculate the mean, median, and mode for the following data set: 6, 8, 10, 10, 15, 15, 15, 20.
9. A company's weekly wages (in hundreds of dollars) for 10 employees are as follows: 25, 30, 30, 35, 35, 35, 40, 45, 45, 50. Calculate the mean, median, and mode. Which measure best represents the typical wage, and why?
10. Discuss the effect of outliers on the mean, median, and mode, with an example.
11. Calculate the weighted mean for a set of test scores where the weights are as follows: Test 1 = 50%, Test 2 = 30%, Test 3 = 20%. Test scores: Test 1 = 80, Test 2 = 75, Test 3 = 90.
12. Describe how the mean, median, and mode can help in understanding the distribution shape of data (e.g., symmetric, skewed).

13. A set of data is heavily skewed to the right. Discuss which measure of central tendency (mean, median, mode) is most appropriate and why.
14. If the mean of a dataset is much higher than the median, what can you infer about the shape of the data distribution?
15. In a classroom, students were given a test, and the mean score was 70. If a new student joins with a score of 95, how will this affect the mean and median?
16. In a given dataset, the mode is significantly lower than both the mean and median. What does this indicate about the data distribution?
17. A real estate agent recorded the prices of houses sold in a neighborhood as follows (in thousands): 150, 160, 200, 250, 1000. Explain which measure of central tendency would be best to represent the "typical" house price.
18. A teacher finds that the mean score in a class test is 65, but the median score is 78. What might this say about the performance of students in the test?
19. A dataset shows that the mode is the best measure of central tendency. Give two examples of real-world scenarios where this might happen.
20. A researcher wants to compare the typical commuting times in two cities. City A's commute times are 25, 30, 35, 40, 100 minutes, and City B's are 20, 22, 23, 25, 28 minutes. Which measure of central tendency should the researcher use for each city, and why?
21. Given the wage distribution table for a factory, calculate:
    (a) the average wage,
    (b) the median wage,
    (c) the wage that appears most frequently,

    Use the following data for your calculations:

Weekly Wage	Number of Workers	Weekly Wage	Number of Workers
Tk. 500-600	10	1000-1100	34
600-700	15	1100-1200	40
700-800	20	1200-1300	25
800-900	26	1300-1400	15
900-1000	30	1400-1500	5

22. Given the following data:

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

    Requirements:
    i. The median and standard deviation.
    ii. The first and third quartiles
    iii. Fourth Decile (D_4) and 85^{th} percentile.

23. In a high school science competition, there are 20 participants: 12 from the junior class and 8 from the senior class. Their scores on a physics quiz are as follows:

    Junior Class: 92, 85, 78, 88, 95, 90, 87, 84, 89, 91, 86, 83
    Senior Class: 77, 79, 82, 88, 90, 76, 84, 80

    Requirements:

    a. Calculate which group has the highest average scores.
    b. Find the median scores of each group.

24. Given below is the frequency distribution of the marks obtained by 90 students. Compute the arithmetic mean, median, and mode:

Marks	No. of students	Marks	No. of students
20-30	2	60-70	18
30-40	12	70-80	10
40-50	15	80-90	9
50-60	20	90-100	4

25. The average weekly wage for 20 persons working in a factory was calculated to be Tk. 2,350. It was later discovered that two were misread as Tk. 1,600 and Tk. 1800 instead of the correct value of Tk. 1,900 and 1700 respectively. Calculate the correct average wage.
26. An investor has a portfolio of three stocks: Stock A, Stock B, and Stock C. He has invested a different amount in each stock, and each stock has a different rate of return. The following table shows the investment amount and rate of return for each stock:

Stock	Investment ($)	Rate of Return (%)
A	5,000	8
B	3,000	5
C	2,000	12

Calculate the weighted average rate of return for the portfolio.

References

Anilkumar, P. (2013). Refining Measure of Central Tendency and Dispersion. *IOSR Journal of Mathematics*, 6, 1-4. https://doi.org/10.9790/5728-0610104.

Chakrabarty, D. (2021). Measuremental Data: Seven Measures of Central Tendency. *International Journal of Electronics and Applied Research.* https://doi.org/10.33665/ijear.2021.v08i01.002.

Meyer, R., Browning, C., & Channell, D. (1995). Expanding Students' Conceptions of the Arithmetic Mean.. *School Science and Mathematics*, 95, 114-117. https://doi.org/10.1111/J.1949-8594.1995.TB15741.X.

Sial, M., & Abid, A. (2023). Measurement of Central Tendencies. *Journal for Research in Applied Sciences and Biotechnology*. https://doi.org/10.55544/jrasb.2.3.29.

MEASURES OF VARIATIONS OR DISPERSION

5

Unit Highlights

- Introduction
- Significance of Measuring Variation
- Key Characteristics of Ideal Measure of Dispersion
- Different Types of Measure of Dispersion

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 5: Measures of Variations or Dispersion

Learning Objectives

By the end of this Unit, you should be able to:

1. Define and explain the significance of measuring variation in understanding data distribution.
2. Calculate the range, inter quartile range (IQR), and their respective coefficients for ungrouped and grouped data.
3. Compute the average deviation and standard deviation for grouped data using both the actual mean and assumed mean methods.
4. Interpret the coefficient of variation and use it to compare relative variability across datasets.
5. Analyze the advantages, limitations, and practical applications of various measures of dispersion, including standard deviation and average deviation.

5.1 Introduction

While measures of central tendency give us a central or typical value in a data set, they don't fully describe the data. Measures of variation, also known as measures of dispersion, provide insight into the spread or variability of data points around the central value. Variation helps us understand how diverse or consistent the data points are. This Unit explores the different ways to quantify variation in a data set.

Class	Scores	Mean
Class X	80, 80, 80, 80, 80, 80, 80, 80, 80, 80	80
Class Y	80, 82, 81, 79, 80, 81, 80, 82, 79, 80	80
Class Z	60, 65, 70, 75, 80, 85, 90, 95, 100, 105	80

Despite each series having the same mean (80), they differ significantly in their distribution or "spread." Class X has no variation, Class Y shows moderate variation, and Class Z has a high level of variation. This shows that while measures of central tendency (like the mean) can provide a central value, they do not describe the full picture of the data's distribution.

Measures of central tendency, like the mean, provide a snapshot of where most values lie, but measures of dispersion reveal the structure and spread of the data. For a comprehensive understanding of any dataset, both types of measures are essential (Lane & Ziemer, 2021; Schacht & Aspelmeier, 2018). For example, in educational contexts, this approach can help teachers make informed decisions about how to address the specific needs of their students, tailoring their methods based on whether students' performances are uniform or varied. Mean tells us the "center" of the data, while dispersion tells us how much the data varies from that center. Together, they offer a fuller, more nuanced view of the data's characteristics and potential implications.

To capture this spread, we use measures of dispersion, such as range, variance, and standard deviation, which help quantify how much the values deviate from the mean.

5.2 Significance of Measuring Variation

Understanding variation is crucial for several reasons:

1. Data Reliability: High variation in data can indicate inconsistency, while low variation implies that data points are closely clustered around the central tendency, suggesting reliability.

2. Comparison Across Data Sets: Variation allows us to compare the spread between different data sets, which is useful when analyzing multiple samples or groups.

3. Risk and Decision-Making: In fields such as finance and manufacturing, measuring variation helps assess risk and make informed decisions. For example, greater variation in returns might indicate higher investment risk.

4. Identifying Outliers: By measuring how spread out the data is, we can identify outliers—data points that are significantly different from the others. Outliers can provide insight or signal issues with data collection.

5.3 Key Characteristics of Ideal Measure of Dispersion

1. The measure should be simple to calculate and understand.
2. It should include every value in the dataset to give an accurate picture of the spread.
3. It should clearly show how much the data varies, with bigger differences making a bigger impact.
4. It should give a specific, positive number, with zero meaning there's no variation at all.
5. The measure should be easy to use in calculations with other statistical tools.
6. It shouldn't be thrown off by very high or low values that are unusual.
7. It should give similar results across different samples from the same population.
8. It should have the same units as the data or be unit-free so it's easy to interpret.

5.4 Different Types of Measure of Dispersion

There are various measures of dispersion:

1. The Range
2. The Interquartile Range (IQR) or Quartile Deviation
3. The Average Deviation
4. The Standard Deviation

The first four methods listed are mathematical, while the Lorenz Curve is a graphical tool. Each method offers unique insights into the distribution and spread of data, helping in different analytical contexts.

5.4.1 The Range

Range is the simplest measure of dispersion, representing the difference between the highest and lowest values in a dataset. It provides a basic sense of how spread out the data is by showing the full span of values. However, it only considers the two extreme points in the data, so it doesn't account for the distribution of values in between.

The range is calculated as:

Range=Maximum Value−Minimum Value

Example

Consider the test scores in a class: 55, 60, 70, 75, 80, 85, 90, and 95.
Maximum score = 95
Minimum score = 55
Range = 95 - 55 = 40. So, the range of these scores is 40.

This measure has a certain appeal because of its simplicity. It is easy to understand and quick to compute, making it ideal for a quick snapshot of the spread in the data. In situations where you need to gauge the general range of values quickly—such as in initial data exploration or when assessing differences between small datasets—the range can be very useful.

However, the range has its limitations, primarily because it only takes into account the two extreme values in the data. If these values are outliers—values that are unusually high or low compared to the rest of the data—the range can be misleading. For instance, imagine a data set of salaries where most employees earn between Tk.50,000 and Tk.70,000, but the CEO takes Tk.1,000,000. The range would be Tk.950,000, which makes it seem like there's massive variability, even though the majority of salaries fall within a much smaller spread.

Moreover, the range doesn't tell us anything about the values between the minimum and maximum. It could be that all values are evenly spaced, or it could be that most values are clustered near the middle with just a couple of extreme values on either end. The range alone doesn't capture any of that nuance, which is why it's often considered an incomplete measure of dispersion.

Because of these limitations, the range is typically used alongside other measures of dispersion, like the interquartile range or standard deviation, which provide a more detailed picture of variability. While the range can quickly reveal the span of data, other measures are needed to understand the consistency or clustering of values within that span. Nonetheless, the range remains a helpful starting point—a simple, direct measure that highlights the extent of variation in a dataset, even if it doesn't tell the full story.

Advantages of Range

1. The range is easy to compute and understand.
2. It gives an immediate sense of the total spread of values.

Limitations of Range

1. Affected by Outliers – Since it only considers the two extreme values, the range can be distorted by outliers (extremely high or low values).
2. Ignores Data Distribution – It doesn't take into account the distribution of values between the minimum and maximum, so it may not reflect the true variability in the data.

Coefficient of Range

The coefficient of range is a relative measure of dispersion that provides insight into the spread of data in relation to its extremes. Unlike the simple range, which only gives the difference between the maximum and minimum values, the coefficient of range standardizes this difference by comparing it to the sum of the maximum and minimum values. This standardization makes it easier to compare variability across different datasets, especially those with different units or scales.

The coefficient of range is calculated using the following formula:

$$\text{Coefficient of Range} = \frac{\text{Maximum Value} - \text{Minimum Value}}{\text{Maximum Value} + \text{Minimum Value}}$$

Let's consider an example to see how the coefficient of range works. Suppose we have two datasets of test scores:

Dataset A: 55, 60, 70, 75, 80, 85, 90, 95
Dataset B: 200, 210, 215, 220, 225, 230, 235, 240

For Dataset A:

Maximum Value = 95
Minimum Value = 55
Range = 95 - 55 = 40

$$\text{Coefficient of Range for Dataset A} = \frac{95 - 55}{95 + 55} = 0.267$$

For Dataset B:

Maximum Value = 240
Minimum Value = 200
Range = 240 - 200 = 40
Coefficient of Range for Dataset B = $\frac{240 - 200}{240 + 200} = 0.091$

Even though both groups have the same absolute range, the coefficient of the range shows us that the first group has relatively more spread or variability (0.267) than the second group (0.091) in relation to its scale. In other words, the values in the first group are more widely spread out relative to their range than those in the second group.

This measure is particularly useful when we are comparing datasets with different units or scales, such as temperatures measured in Fahrenheit and Celsius or prices in two different currencies. By using the coefficient of range, we get a standardized measure of variability that allows us to make fair comparisons without worrying about the units or scale of the data.

However, the coefficient of range has its limitations. Since it only considers the highest and lowest values, it can be sensitive to outliers. For example, if there's one unusually high or low value, it could skew the coefficient, making the dataset seem more or less spread out than it is. Also, the coefficient of range doesn't tell us anything about how values are distributed within the range. It is possible to have a high coefficient of range even if most of the data points are clustered closely together, with only a few values stretching the range.

In practice, the coefficient of range is often used as a starting point for understanding variability in a dataset, particularly in the early stages of data analysis or when making quick comparisons between groups. For a more detailed view of how data is spread out, it is common to look at other measures like standard deviation or the interquartile range, which take into account more of the data points and provide a fuller picture of variability.

Illustration: The following grouped data represents the heights (in cm) of 50 students:

Height Range (Class Interval)	Frequency
140–149	5
150–159	8
160–169	12
170–179	15
180–189	10

Find the Range and coefficient of the Range.

Solution:

The range is calculated as:
Range=Maximum Value−Minimum Value
The maximum value is the upper boundary of the highest class interval: 189.
The minimum value is the lower boundary of the lowest class interval: 140.
So, Range=189−140=49
The coefficient of range is calculated as:

Coefficient of Range = $\frac{\text{Maximum Value} - \text{Minimum Value}}{\text{Maximum Value} + \text{Minimum Value}}$

Substitute the values:

Coefficient of Range = $\frac{189-140}{189+140}$= 0.149

The range of 49 cm indicates the total spread of the students' heights. The coefficient of range, approximately 0.149, shows that the relative dispersion is moderate, suggesting a fairly consistent set of height measurements.

5.4.2 The Interquartile Range (IQR) or Quartile Deviation

The quartile deviation, also called the interquartile range, is a statistical tool that helps us understand how spread out the middle values of a dataset are, focusing on the central 50% of data points. It is particularly helpful when we want to get a sense of the variability around the median, without being overly influenced by extreme high or low values. This makes the quartile deviation a reliable measure of dispersion, especially for data that isn't evenly distributed or is skewed, with outliers that could distort our understanding of the spread.

To understand quartile deviation, it helps to start with quartiles. Quartiles divide a dataset into four parts:

Q_1 (First Quartile), marking the value below which the lowest 25% of data falls.

Q_2 (Median), which is the middle value, with 50% of the data below it.

Q_3 (Third Quartile), showing the point below which 75% of the data lies.

The Interquartile Range (IQR), calculated as $Q_3 - Q_1$, represents the spread of the middle 50% of values in the dataset. It is a measure of variability that indicates how concentrated the middle values are.

The Quartile Deviation, or Interquartile Range, is half of the Interquartile Range:

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

This value represents the average distance of the middle 50% of data points from the median, giving us a sense of how spread out these values are.

Suppose, we have the following dataset representing exam scores in an ascending format:

45, 50, 55, 60, 65, 70, 75, 80, 85.

Here, Q_1 (25th percentile) = 55 (the median of the lower half of the data).

Q_3 (75th percentile) = 75 (the median of the upper half of the data).

IQR=$Q_3 - Q_1$=75−55=20

Quartile Deviation = $\frac{20}{2}$ = 10

So, the quartile deviation for this dataset is 10, meaning that the middle 50% of the scores are, on average, 10 points away from the median.

The quartile deviation has some clear advantages. Because it looks only at the middle 50% of the data, it's less sensitive to extreme values—those scores that are unusually high or low. In other words, it's resistant to the influence of outliers, which can distort other measures like the full range. For this reason, it is especially useful in skewed datasets, like income distributions, where a few very high values can make it look like there is more variability than there really is.

However, the quartile deviation also has its limitations. Since it only considers the middle 50%, it ignores the upper and lower quarters of the data, potentially missing important

variability in those areas. For smaller datasets, where quartiles are less stable, the quartile deviation may also be less meaningful.

In practice, the quartile deviation is a useful measure in fields where data might be skewed or contain outliers. For example, in analyzing home prices in a neighborhood, the quartile deviation can give a more stable view of typical price variability, without being skewed by a few very expensive homes.

Coefficient of Quartile Deviation

The coefficient of quartile deviation, also known as the coefficient of quartile variation, is a measure of relative dispersion based on the interquartile range. It helps assess the spread of the middle 50% of values in a data set and is often used when the data is skewed or when focusing on the central tendency rather than extreme values.

The formula for the coefficient of quartile deviation is:

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q3+Q1}$

For the above example, the coefficient of quartile deviation would be,

Coefficient of Quartile Deviation = $\frac{75-55}{75+55} = \frac{20}{130} = 0.15$

The resulting coefficient of quartile deviation gives us insight into the variability of the middle data without letting extreme values unduly influence our assessment. A higher coefficient suggests that the values around the median are more spread out, whereas a lower coefficient indicates that the values are closer together. For example, in income data, where extreme values might pull a standard deviation measure far from the center, the coefficient of quartile deviation offers a stable way to understand variability around the median income level without the distortion of exceptionally high or low incomes.

The quartile deviation and coefficient of quartile deviation of grouped data are shown in the following illustration.

Illustration: The following data represents the monthly household income of 50 families in a neighborhood. The data is grouped into class intervals as follows:

Income Range (in thousands of Taka)	Frequency
0 – 10	3
10 – 20	5
20 – 30	8
30 – 40	14
40 – 50	10
50 – 60	6
60 – 70	4
Total	50

Requirements: Calculate i. interquartile range, ii. quartile deviation, and iii. coefficient of quartile deviation.

Answer:

We'll first calculate the cumulative frequency for each class.

Income Range (in thousands of Taka)	Midpoints (X)	Frequency (f)	Cumulative Frequency (c.f)
0 – 10	5	3	3
10 – 20	15	5	8
20 – 30	25	8	16
30 – 40	35	14	30
40 – 50	45	10	40
50 – 60	55	6	44
60 – 70	65	4	50

The total frequency N is 50. To find the position of the Q_1, Q_1 lies on $\frac{50}{4}^{th}$ observation, 12.5^{th} observation. The class interval containing the 12.5th data point is the 20-30 income range because 12.5 lies between the cumulative frequency of 8 and 16. So, Q_1 lies in the 20-30 range.

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

L= 20

$p.c.f.$ = 8 (cumulative frequency before the first quartile Q_1 class).

f = 8 (frequency of the first quartile Q_1 class).

i = 10 (class width).

$Q_1 = 20 + \frac{\frac{50}{4} - 8}{8} \times 10$

= 20+5.625

= 25.625 thousand Taka.

Q_3 lies on $\frac{3*50}{4}^{th}$ observation, 37.5^{th} observation. The class interval containing the 37.5th data point is the 40-50 income range because 37.5 lies between the cumulative frequency of 30 and 40. So, Q_3 lies in the 40-50 range.

$Q_3 = L + \frac{\frac{3N}{4} - p.c.f}{f} \times i$

L = 40

pcf = 30 (cumulative frequency before the third quartile Q_3 class).

f = 10 (frequency of the third quartile Q_3 class).

i = 10 (class width).

$Q_3 = 40 + \frac{\frac{3*50}{4} - 30}{10} \times 10$

= 40 + 7.5

= 47.5 thousand Taka.

i. The Interquartile Range (IQR) by subtracting Q_1 from Q_3:

$IQR = Q_3 - Q_1 = 47.5 - 25.625 = 23.875$

The Interquartile Range (IQR) of the household income data is 23.88 thousand Taka. This means that the middle 50% of the data (household incomes) lie between Tk. 25,625 and Tk. 47,500 per month.

ii. Quartile Deviation (QD) = $\frac{Q_3-Q_1}{2} = \frac{23.875}{2}$ = 11.9375 thousand Taka.

The Quartile Deviation (QD) gives an absolute measure of the spread of the middle 50% of the data. In this case, the middle 50% of the income data lies within 11.94 thousand Taka of each other.

iii. Coefficient of Quartile Deviation (CQD) = $\frac{Q_3-Q_1}{Q_3+Q_1} = \frac{23.875}{73.125} = 0.326$

The Coefficient of Quartile Deviation (CQD) gives a relative measure of the spread, allowing to compare the dispersion between datasets with different units or scales. In this case, the value of 0.326 means that the Quartile Deviation is about 32.6% of the sum of the first and third quartiles.

5.4.3 The Average Deviation

The average deviation, or mean absolute deviation (MAD), is a straightforward way to measure how much values in a dataset vary from the center. For example, imagine a set of test scores, and the goal is to understand how close each score is to the overall average. The average deviation calculates the average distance of each data point from the central value, typically the mean.

To find the average deviation, a central value is chosen, usually the mean (or sometimes the median). Then, for each data point, the distance from this central value is measured. Rather than letting positive and negative distances cancel each other out, the absolute (positive) value of each distance is taken, so only the amount each point differs from the center is measured. Adding up all these absolute values and dividing by the total number of data points yields the average deviation, which shows the typical distance of each data point from the center.

The formula for average deviation when using the mean as the central value is:

$$\text{Average Deviation}_{(Mean)} = \frac{\sum|\mathbf{X}-\overline{X}|}{\mathrm{N}}$$

$$\text{Average Deviation}_{(Median)} = \frac{\sum|\mathbf{X}-Me\ |}{\mathrm{N}}$$

where:

X represents each data point,

$\overline{X}$ is the mean of the dataset,

N is the number of data points.

Let's say we have a small dataset of exam scores: 50, 55, 60, 65, and 70.

Here, Mean, $\overline{X} = \frac{50+55+60+65+7}{5} = 60$

Average Deviation = $\frac{|50-60|+|55-60|+|60-60|+|65-6\ |+|70-60|}{5} = \frac{10+5+0+5+}{5} = 6$

The beauty of the average deviation lies in its simplicity—it is easy to calculate and interpret, giving a straightforward view of how spread out the values are around the center. If the average deviation is low, it means the values are tightly clustered around the mean, indicating less variability. A higher average deviation shows that the data points are more spread out, signifying greater variability.

However, there are some things to keep in mind. The average deviation is not as commonly used as other measures like the standard deviation because it's a bit simpler and does not fit as well into certain statistical models, especially those that assume a normal distribution. But it is an excellent choice when you need a clear, easy-to-understand measure of spread, and it is particularly helpful for getting a quick sense of how much data points differ from the center without worrying about extreme values (outliers).

Advantages

1. **Simplicity** – It is straightforward to calculate and interpret.

2. **Realistic Representation** – By using absolute values, it provides a clear picture of variability without canceling out positive and negative deviations.

Limitations

1. **Less Common in Advanced Analysis** – Unlike standard deviation, average deviation is not used as frequently in advanced statistical analyses, partly because it does not fit as neatly with statistical models that assume a normal distribution.

2. **Less Sensitive to Outliers** – While this can be an advantage in certain cases, it may also make average deviation less effective for detecting extreme variability than standard deviation.

Coefficient of Average Deviation

The Coefficient of Average Deviation is a useful tool for understanding how spread-out data values are in relation to a central point, like the mean or median. Imagine we are looking at a set of numbers, such as prices, incomes, or test scores, and we want to understand not just the overall average spread, but how this spread compares to the central value itself. The coefficient of average deviation answers this by expressing the average deviation as a proportion of the central value, giving us a percentage or relative measure of dispersion. The coefficient of average deviation is calculated as follows:

$$\text{Coefficient of Average Deviation} = \frac{\text{Average Deviation}}{\text{Mean or Median}}$$

where:

-The Average Deviation (or mean absolute deviation) is the average of the absolute deviations from the central value (mean or median).

-The denominator, Mean or Median, is the chosen central value of the dataset.

For the above example, the Coefficient of Average Deviation would be

$$\text{Coefficient of Average Deviation} = \frac{\text{Average Deviation}}{\text{Mean or Median}} = \frac{6}{60} = 0.10 \text{ or } 10\%$$

This result means that, on average, each value is about 10% away from the central value, giving a clear picture of relative variability.

By showing the spread as a fraction of the mean or median, the coefficient of average deviation allows us to compare variability between different datasets or measures, even if the units are different. A lower coefficient means that values are closely clustered around the center, while a higher coefficient shows a greater spread. This measure is especially handy when working with skewed data, where extreme values might distort other measures of spread, but the average deviation keeps the focus on typical variability.

Illustration:

Calculate the average deviation and coefficient of average deviation for the heights (in cm) of students in two different classes.

Class X Heights (cm)	Class Y Heights (cm)
150	140
155	145
160	150
165	155
170	160
	165

Solution:

Class X: Mean, $\overline{X}=\frac{150+155+160+}{5}$ = 160

Class Y: Mean, $\overline{Y}=\frac{140+145+150+155+160+165}{6}$ = 152.5

Class X: Average Deviation= $\frac{|150-160|+|155-160|+|160-160|+|165-160|+|170-160|}{5}=\frac{10+5+0+5+10}{5}=6$

Class Y: Average Deviation= $\frac{|140-152.5|+|145-152.5|+|150-152.5|+|155-152.5|+|160-152.5|+|165-152.5|}{6}=\frac{12.5+7.5+2.5+2.5+7.5+12.5}{6}=7.5$

The Coefficient of Average Deviation = $\frac{\text{Average Deviation}}{\text{Mean or Median}}$

Class X: Coefficient of Average Deviation = $\frac{6}{160}$ = 0.0375 or 3.75%

Class Y: Coefficient of Average Deviation = $\frac{7.5}{152.5}$ = 0.0492 or 4.92%

Calculation of Average Deviation for grouped data

The formula for calculating the average deviation from the mean or the median for grouped data is:

$\text{A. D.}_{(Mean)}=\frac{\sum f|\text{X}-\overline{X}|}{\text{N}}$ or $\text{A. D.}_{(Med.)}=\frac{\sum f|\text{X}-\text{Median}|}{\text{N}}$

where:

X represents each data point,
f = frequency of each class interval,
$\overline{X}$ is the mean of the dataset,
Med is the median of the dataset,
N is the number of data points.

Illustration: Calculate the average deviation from the median for the following grouped data representing the time (in minutes) taken by different batches of students to complete a test:

Time Taken (minutes)	Frequency (No. of Students)
10–20	5
20–30	8
30–40	12
40–50	7
50–60	3

Solution:

Total number of students, N=5+8+12+7+3=35.

The median position is $\frac{N}{2}=\frac{35}{2}$=17.5, we look for the class interval that contains the 17.5th observation. So, the median class is 30-40 because 17.5th lies between 13 and 25.

Median = L + $\frac{\frac{N}{2}-p.c.f}{f}\times i$ = 30 + $\frac{17.5-13}{12}\times 10$ = 30+3.75 = 33.75

Time Taken (minutes)	Midpoint X	Frequency (No. of Students)f	Cumulative Frequency	|X−Median|	f.|X−Median|
10–20	15	5	5	|15−33.75|=18.75	93.75
20–30	25	8	13	|25−33.75|=8.75	70.00
30–40	35	12	25	|35−33.75|=1.25	15.00
40–50	45	7	32	|45−33.75|=11.25	78.75
50–60	55	3	35	|55−33.75|=21.25	63.75
		N=35			$\sum f$|X−Median|=321.25

Average Deviation, A. D.$_{(Med.)}$ = $\frac{321.25}{35}$ = 9.18

The average deviation from the median is 9.18 minutes.

Coefficient of Average Deviation = $\frac{A.D.}{Median}\times 100=\frac{9.18}{33.75}\times 100$ = 27.2%

The Coefficient of Average Deviation is approximately 27.2%. This means that the average deviation is 27.2% of the median, indicating the relative variability of the dataset around the median.

5.4.4 Standard Deviation

The Standard Deviation (SD) is one of the most commonly used measures of dispersion or variability in a dataset. It helps quantify the amount of variation or spread around the mean (average) value of the data. Understanding standard deviation provides insight into the consistency and reliability of data.

The concept was introduced by Karl Pearson in 1893, and it quickly became one of the most popular ways to measure variation. Earlier measures of variation had limitations, but standard deviation overcomes many of them. Looking at the spread of values gives us a sense of whether most numbers are close to the mean (average) or if they are more widely dispersed (Jeong & Chong, 2019).

Standard deviation offers a way to measure how "spread out" the values in a dataset are. When data points are close to the mean, the standard deviation is small, indicating low variability. Conversely, a large standard deviation indicates that data points are spread out over a wider range, showing higher variability. This can help in comparing different datasets, identifying outliers, or making inferences about populations in statistics.

For example, consider the average temperatures of two cities. If one city has a standard deviation of 2 degrees and the other has 8 degrees, we can say the second city has more variability in temperature, meaning it experiences more fluctuation from day to day.

To calculate standard deviation:

For a sample dataset, the formula for standard deviation (s) is:

$$s = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$$

For a population dataset, the formula is:

$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

Where:
X = Each individual data point
$\overline{X}$ = Sample mean
μ = Population mean
N = Total number of data points in the population
n = Total number of data points in the sample

The standard deviation is based on the concept of deviation from the mean. It involves calculating how much each data point differs from the mean and then averaging these differences in a way that accounts for all deviations. To calculate standard deviation:

1. Calculate the Mean ($\overline{X}$ or μ pronounced as mu): This is the central point or the average of all data points in the dataset.

2. Find Deviations from the Mean: For each data point, subtract the mean to find the "deviation" or difference between the data point and the mean ($X - \overline{X}$). Deviations can be positive (above the mean) or negative (below the mean). e.g., $X_1 - \overline{X}, X_2 - \overline{X}, \ldots.., X_n - \overline{X}$

3. Square Each Deviation: Squaring each deviation $(X - \overline{X})^2$ makes all values positive and gives more weight to larger deviations, thus emphasizing data points that are farther from the mean. e.g., $(X_1 - \overline{X})^2, (X_2 - \overline{X})^2, \ldots\ldots, (X_n - \overline{X})^2$

4. Calculate the Average of the Squared Deviations (Variance): For a sample dataset, divide the sum of squared deviations by n−1 (where n is the sample size) to get the sample variance.

For a population dataset, divide by N to get the population variance.

5. Take the Square Root of the Variance: This step brings the measure back to the original units of the data, providing the standard deviation.

Let's use a simple dataset of test scores: 40, 50, 60, 70, and 80.

$\overline{X} = \frac{40+50+60+70+8}{5} = 60$

Variance, $s^2 = \frac{\sum(X-\overline{X})^2}{n-1} = \frac{(40-60)^2+(50-60)^2+(60-6\ )^2+(70-60)^2+(80-60)^2}{5-1} = \frac{400+100+0+100+400}{5-1} = 250$

Standard Deviation**:**

$$s = \sqrt{250} \approx 15.81$$

The Standard Deviation of the dataset is approximately 15.81. This means that, on average, each test score varies from the mean score of 60 by about 15.81 points. Standard deviation gives a clearer picture of data variability and is commonly used in many fields for statistical analysis.

Interpretation of Standard Deviation

A small standard deviation indicates that data points are close to the mean, showing low variability. A large standard deviation indicates that data points are spread out from the mean, showing high variability. For example, if the standard deviation of heights in a class is small, most students are of similar height. A large standard deviation would mean a wider range of heights.

Practical Uses of Standard Deviation

1. **Comparing Datasets**: If two datasets have different means, the standard deviation allows you to compare their spread or variability. For instance, it can help compare consistency in test scores between two different classes.
2. **Quality Control**: In manufacturing, standard deviation measures product consistency. A low standard deviation means products are consistent and meet quality standards.
3. **Risk Assessment**: The standard deviation measures the volatility of stock prices or investment returns in finance. A higher standard deviation implies higher risk and potential return.
3. **Research and Statistics**: Standard deviation is fundamental in statistical analyses, especially when determining confidence intervals, hypothesis testing, and normal distribution.

Limitations of Standard Deviation

1. **Sensitive to Outliers**: Standard deviation can be affected by extreme values or outliers, as these significantly increase the squared deviations. This may give an inflated sense of variability if there are outliers in the data.
2. **Assumes a Normal Distribution**: Standard deviation is most meaningful for data that is symmetrically distributed. For skewed or non-normal data, other measures like the interquartile range may be more informative.
3. **Not Always Intuitive**: For non-technical users, interpreting the meaning of standard deviation in real-world terms can be difficult.

Steps to Calculate Standard Deviation for Grouped Data

For grouped frequency distribution, the standard deviation can be calculated using two main methods:

a. Standard Deviation by Actual Mean Method

1. **Find the Midpoint (Class Mark) for Each Class Interval**: The midpoint X_i of each interval is calculated by taking the average of the lower and upper limits of each class interval.
2. **Calculate the Mean $\overline{X}$ Using the Midpoints**: Multiply each midpoint X_i by the corresponding frequency f to get fX_i. Then, sum up all fX_i values and divide by the total frequency (N).
3. **Find Deviation for Each Class Interval**: Subtract the mean from each midpoint to get the deviation $(X_i - \overline{X})$. Then, square each deviation to get $(X_i - \overline{X})^2$
4. **Multiply the Squared Deviations by Frequency**: Multiply $(X_i - \overline{X})^2$ by the corresponding frequency f to get $f(X_i - \overline{X})^2$
5. **Calculate Variance and Standard Deviation**: Sum up all $f(X_n - \overline{X})^2$ values to get $\sum f(X_n - \overline{X})^2$. Then, divide this sum by the total frequency N to get the variance. Finally, take the square root of the variance to get the standard deviation.

$$\sigma = \sqrt{\frac{\sum f(X_i - \overline{X})^2}{N}}$$

Illustration: Suppose we have the following data for weekly hours of exercise among students. Find the mean and standard deviation for grouped data.

Weekly Hours (Class Interval)	Frequency (f)
0 – 5	4
5 – 10	6
10 – 15	10
15 – 20	8
20 – 25	2

Solution:

Scores (X)	Midpoint X_i	Frequency f	fX_i	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	$f(X_i - \overline{X})^2$
0 – 5	2.5	4	10	-9.67	93.51	374.04
5 – 10	7.5	6	45	-4.67	21.81	130.86
10 – 15	12.5	10	125	0.33	0.11	1.1
15 – 20	17.5	8	140	5.33	28.41	227.28
20 – 25	22.5	2	45	10.33	106.73	213.46
Total		N = 30	fX_i =365			$\sum f(X_i - \overline{X})^2$ = 946.74

$\overline{X} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{365}{30} = 12.17$

The standard deviation, $\sigma = \sqrt{\frac{\sum f(X_i - \overline{X})^2}{N}} = \sqrt{\frac{946.74}{30}} = \sqrt{31.56} \approx 5.62$

b. Standard Deviation by Assumed Mean Method

This method simplifies calculations, especially when dealing with large numbers, by using an assumed mean (a convenient central value close to the middle of the data). The steps are:

1. Selecting an assumed mean *A*, usually close to the midpoint of the data range.
2. Finding deviations $d = \frac{X_i - A}{i}$ for each class interval.
3. Multiplying each deviation by its frequency to find fd.
4. Summing fd values and dividing by *N* to adjust the assumed mean, if necessary.
5. Calculating d^2 for each class, then, finding fd^2 and summing fd^2 values.
6. Using the formula for variance with adjustments based on the assumed mean, and then taking the square root to obtain the standard deviation.

Formula:

The formula for standard deviation when deviations are taken from an assumed mean A is:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - (\frac{\sum fd}{N})^2} \times i$$

where:
$d = X_i - A$ (deviation of each class midpoint from the assumed mean)
A = assumed mean (a central value close to the middle of the data range)
i = class interval
f = frequency of each class interval
N = total frequency

Illustration:

Suppose we have the following data for students' heights in a class. Find the mean and the standard deviation for grouped data using the assumed mean.

Height (cm)	Frequency (f)
150 – 155	5
155 – 160	8
160 – 165	15
165 – 170	16
170 – 175	6

Solution:

Let's use 162.5 as the assumed mean since it's close to the center.

Height (cm)	Frequency (f)	Midpoint X_i	$d = \frac{X_i - 162.5}{5}$	d^2	fd	fd^2
150 – 155	5	152.5	-2	4	-10	20
155 – 160	8	157.5	-1	1	-8	8
160 – 165	15	162.5	0	0	0	0
165 – 170	16	167.5	1	1	16	16
170 – 175	6	172.5	2	4	12	24
	N= 50				$\sum fd$=10	$\sum fd^2$=68

Mean $\overline{X} = A + \frac{\sum fd}{N} \times i$= 162.5+$\frac{10}{50}$×5 =*163.5*

Standard deviation, $\sigma = \sqrt{\frac{\sum fd^2}{N} - (\frac{\sum fd}{N})^2} \times i$

$= \sqrt{\frac{68}{50} - (\frac{10}{50})^2} \times 5$

$= \sqrt{1.36 - (.20)^2} \times 5$

$= \sqrt{1.32} \times 5 = 1.1489 \times 5 = 5.74$

The standard deviation for this grouped data is approximately 5.74 cm**.**

Mathematical Properties of Standard Deviation

1. **Non-Negativity:** The standard deviation is always non-negative:

$$s \geq 0$$

This is because it represents a distance (or spread) around the mean, so it cannot be negative. A standard deviation of zero implies that all values in the dataset are the same, with no variation.

2. Affected by the Scale of Data

If all values in a dataset are multiplied by a constant *k*, the standard deviation is also multiplied by *k*:

$s(k \cdot X) = |k| \cdot s(X)$

For example, if a dataset has a standard deviation of 5 and all values are doubled, the new standard deviation becomes 2×5=10. This property shows that standard deviation scales with the unit of measurement.

3. Shift Property (Unaffected by Additive Constants)

Adding or subtracting a constant c to every value in a dataset does not affect the standard deviation:

$s(X+c)=s(X)$

For instance, if you add 10 to each value in a dataset, the standard deviation remains the same. This is because the measure of dispersion or spread among the values doesn't change with an additive constant.

4. Relation to Variance

Standard deviation is the square root of variance:

$s=\sqrt{\mathrm{Var}(X)}$

Variance (denoted σ^2 or s^2) measures the average squared deviation from the mean, and standard deviation brings it back to the original units of the data by taking the square root.

5. Minimum Property

The standard deviation measures the spread of data from its mean, and it is minimized when the mean is chosen as the central point. No other point (e.g., median, mode) provides a smaller sum of squared deviations. Mathematically, for any constant c:

$$\sum\left(X_i-\overline{X}\right)^2 \leq \sum(X_i-c)^2$$

This property shows why the mean is the best central measure for standard deviation.

6. Combining Standard Deviations of Groups (For Independent Groups)

When combining two or more independent groups, the combined standard deviation can be calculated based on the group sizes, means, and individual standard deviations. If we have two groups A and B with sizes n_A and n_B, means $\overline{X}_A$ and $\overline{X}_B$, and standard deviations s_A and s_B, the combined sample standard deviation s_{combined} is:

$$s_{\text{combined}}=\sqrt{\frac{(n_A-1)s_A^2+(n_B-1)s_B^2+n_A\left(\overline{X}_A-\overline{X}_{combined}\right)^2+n_B\left(\overline{X}_B-\overline{X}_{combined}\right)^2}{n_A+n_B-1}}$$

where $\overline{X}_{combined}$ is the combined mean, $\overline{X}_{combined}=\frac{n_A\overline{X}_A+n_B\overline{X}_B}{n_A+n_B}$

Combined population standard deviation

$$\sigma_{\text{combined}}=\sqrt{\frac{N_A\sigma_A^2+N_B\sigma_B^2+N_A(\mu_A-\mu_{combined})^2+N_B(\mu_B-\mu_{combined})^2}{N_A+N_B}}$$

where $\overline{X}_{combined}$ is the combined mean. $\mu_{combined}=\frac{N_A\mu_A+N_B\mu_B}{N_A+N_B}$

7. Sensitivity to Outliers

Standard deviation is sensitive to outliers. Since it is based on squared deviations, large deviations have a disproportionately large impact on the standard deviation. This means that if there are extreme values in a dataset, the standard deviation can be quite large, reflecting the increased spread.

8. Connection to Normal Distribution

In a normal distribution**:**
Approximately **68%** of data lies within one standard deviation of the mean.
About **95%** falls within two standard deviations.
Around **99.7%** is within three standard deviations.

This property, known as the **68-95-99.7 rule**, allows us to predict data distribution. For example, if students' test scores follow a normal distribution with a mean of 76 and an SD of 11.94, we can expect about 68% of students to score between 64.06 (mean - SD) and 87.94 (mean + SD).

9. Standard Deviation of Sample vs. Population

For sample data, the formula for standard deviation differs slightly from the population standard deviation. For a sample s is given by:

$$s = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n-1}}$$

while for a population it's:

$$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$$

The n−1 denominator in the sample standard deviation is known as Bessel's correction and corrects for bias in estimating the population standard deviation.

Coefficient of Variation

The coefficient of variation (CV) is a standardized measure of the dispersion of data points in a data series around the mean. Unlike other measures of spread, such as standard deviation, the coefficient of variation expresses the extent of variability in relation to the mean of the population. This makes it especially useful for comparing the relative variability of datasets that have different units or vastly different means.

The CV is particularly useful in fields like finance and science, where it is essential to compare the degree of variation between datasets of different magnitudes or units. For instance, in investment analysis, the CV helps compare the risk (standard deviation) relative to the expected return (mean), making it easier to decide between different investment options.

The coefficient of variation is calculated as:

$CV=(\frac{\sigma}{\overline{X}})\times 100$

where:

σ is the standard deviation of the dataset

$\overline{X}$ is the mean of the dataset.

Expressing it as a percentage by multiplying by 100 gives an intuitive sense of the variability in terms of a proportion of the mean.

Illustration:

Suppose two different factories, Factory A and Factory B, have the following weekly wage data for their workers:

Factory	Average Weekly Wages (Taka)	Standard Deviation of Wages (Taka)
A	4500	500
B	4100	300

Find the coefficient of variation for each factory.

Solution:

For Factory A**:**

$CV_A=(\frac{\sigma_A}{\overline{X}_A})\times100=(\frac{500}{4500})\times100=11.11\%$

For Factory B**:**

$CV_B=(\frac{\sigma_B}{\overline{X}_B})\times100=(\frac{300}{4100})\times100=7.32\%$

Factory A has a CV of 11.11%, while Factory B has a CV of 7.32%.This suggests that Factory A has greater relative variability in wages compared to Factory B, even though Factory A has a higher average wage.

Illustration: In two departments, X and Y, within a company, the average monthly working hours and standard deviation of hours are provided below:

Department	Average Monthly Hours	SD of Hours	No. of Employees
X	160	15	150
Y	145	10	100

Requirements:

1. Which department, X or Y, has a higher average number of monthly working hours?
2. Which department shows greater variability in the number of monthly working hours?
3. What is the combined mean and standard deviation of the monthly working hours for all employees in both departments?

Solution:

1. Comparing Average Monthly Hours: Department X has a higher average monthly working hours (160 hours) than Department Y (145 hours).
2. To find the department with greater variability, we calculate the Coefficient of Variation (CV) for each department.

   The formula for the coefficient of variation is:

   $CV=(\frac{\sigma}{\overline{X}})\times100$

   For Department X:

   $CV_X=(\frac{\sigma_X}{\overline{X}_X})\times100=(\frac{15}{160})\times100=9.375\%$

   For Department Y:

   $CV_Y=(\frac{\sigma_Y}{\overline{X}_Y})\times100=(\frac{10}{145})\times100=6.90\%$

   Department X shows greater variability in the number of monthly working hours because its coefficient of variation (9.375%) is higher than that of Department Y (6.90%).

3. Combined Mean:

   $\overline{X}_{combined}=\frac{N_A\overline{X}_A+N_B\overline{X}_B}{N_A+N_B}=\frac{(150\times160)+(100\times145)}{150+100}=\frac{24000+14500}{250}=154$ hours

   Combined Standard Deviation:

   $\sigma_{combined}=\sqrt{\frac{N_A\sigma_A^2+N_B\sigma_B^2+N_A(\overline{X}_A-\overline{X}_{combined})^2+N_B(\overline{X}_B-\overline{X}_{combined})^2}{N_A+N_B}}$

   $=\sqrt{\frac{150(15)^2+100(10)^2+150(160-154)^2+100(145-154)^2}{150+100}}=\sqrt{\frac{150\times225+100\times100+150\times36+100\times81}{250}}$

   $\sqrt{\frac{33750+10000+5400+8100}{250}}=\sqrt{\frac{57250}{250}}=15.13$ hours

Review Questions

1. What is the purpose of using measures of dispersion in data analysis?
2. Explain the difference between range, variance, and standard deviation.
3. How is the range calculated, and what are its limitations as a measure of dispersion?
4. Define variance and explain how it differs for a sample and a population.
5. What is standard deviation, and why is it preferred over variance in many cases?
6. What is the coefficient of variation, and how does it help compare the relative dispersion of different datasets?
7. Describe the difference between absolute and relative measures of dispersion. Provide examples.
8. What is interquartile range (IQR), and why is it a useful measure of dispersion, especially for skewed data?
9. How do outliers affect different measures of dispersion like range, variance, and standard deviation?
10. Two datasets are given:
    Dataset A: 10, 12, 14, 16, 18
    Dataset B: 5, 15, 25, 35, 45
    Compare the range of these datasets and explain which dataset shows more variation.
11. The weekly income (in dollars) of 30 individuals is grouped as follows:

Income Range	Frequency
200–400	6
400–600	10
600–800	8
800–1000	4
1000–1200	2

    Requirements: a. Calculate the range of the data.
    b. Calculate the coefficient of range**.**
12. The following data shows the weights (in kg) of 10 people:
    50, 55, 60, 65, 70, 75, 80, 85, 90, 95
    Requirements: a. Calculate the Interquartile Range (IQR).
    b. Calculate the Coefficient of Quartile Deviation**.**
13. The daily wages (in Taka) of workers in a factory are given below:

Wage Range	Frequency
100–200	5
200–300	8
300–400	15
400–500	12
500–600	10

    Calculate the IQR and Coefficient of Quartile Deviation.
14. The marks of 5 students in a test are: 20, 25, 30, 35, 40.
    a. Calculate the Average Deviation about the mean.
    b. Calculate the Coefficient of Average Deviation**.**
15. The ages of individuals in a group are as follows:

Age Range	Frequency
20–30	4
30–40	6
40–50	10
50–60	8
60–70	2

    Requirements: a. Calculate the Average Deviation about the mean.
    b. Calculate the Coefficient of Average Deviation.

16. The scores of a student in five subjects are: 45, 50, 55, 60, 65. Calculate the Standard Deviation and Calculate the Coefficient of Variation (CV).
17. The monthly expenditures (in dollars) of families in a locality are given below:

Expenditure Range	Frequency
1000–2000	6
2000–3000	10
3000–4000	15
4000–5000	12
5000–6000	7

Requirements:
   a. Calculate the Standard Deviation using the Actual Mean Method.
   b. Calculate the Coefficient of Variation (CV).

18. The following data shows the number of hours spent studying by a group of students:

Hours Studied	Frequency
0–2	4
2–4	6
4–6	8
6–8	10
8–10	2

Calculate the Standard Deviationusing the Actual Mean Method**.**

19. The following data shows the weights (in kg) of a group of individuals:

Weight Range (Class Interval)	Frequency (f)
50–59	5
60–69	8
70–79	10
80–89	6
90–99	4

Calculate the Standard Deviation using the Assumed Mean Method.

20. Two datasets represent the performance of two students in 5 tests.

Student A: 70, 75, 80, 85, 90
Student B: 60, 65, 70, 75, 80

Calculate the Coefficient of Variation (CV) for both students and determine which student has more consistent performance.

References

Jeong, J., & Chong, S. (2019). Adaptation to mean and variance: Interrelationships between mean and variance representations in orientation perception. *Vision Research*, 167, 46-53. https://doi.org/10.1167/19.10.192c.

Lane, D., & Ziemer, H. (2021). Central Tendency. *Encyclopedic Dictionary of Archaeology*. https://doi.org/10.4135/9780857020024.n11.

Schacht, S., &Aspelmeier, J. (2018). Measures of Variability. *Statistical Applications for the Behavioral and Social Sciences*. https://doi.org/10.4324/9780429497308-5.

SKEWNESS, MOMENTS AND KURTOSIS

6

Unit Highlights

- Introduction
- Skewness
- Moments
- Kurtosis

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 6: Skewness, Moments and Kurtosis

Learning Objectives

By the end of this Unit, you should be able to:

1. Define and calculate skewness, moments, and kurtosis, and understand their role in describing data distributions.
2. Interpret skewness to determine the symmetry of data and apply moments to assess central tendency, spread, and shape.
3. Understand the significance of kurtosis in identifying outliers and tail behavior in data.

6.1 Introduction

In business statistics, understanding the shape and characteristics of a data distribution is crucial for making informed decisions. While measures like the mean and variance provide essential insights into central tendency and spread, skewness and kurtosis offer additional information about the distribution's symmetry, shape, and outliers. These concepts help analysts and decision-makers assess the behavior of business data, especially when dealing with large datasets.

6.2 Skewness

Skewness in statistics refers to the lack of symmetry in a distribution. A distribution is symmetric when it is evenly balanced around its central point. Skewness measures the direction and degree of asymmetry in the distribution, giving insight into whether the data tends to cluster more on one side of the mean.

6.2.1 Types of Skewness

There are two types of skewness:

Positive Skew (Right Skewed): In a positively skewed distribution, the right tail (the larger values) is longer than the left tail (the smaller values). This implies that most data points are concentrated on the lower end of the scale, and a few large values stretch the distribution to the right.

Characteristics:

- **Mean > Median > Mode**.
- The mean is pulled towards the right due to the few larger values in the distribution.

Example: Income distribution, where most people earn average or lower incomes, but there are a few very high-income earners who skew the distribution to the right.

Negative Skew (Left Skewed): In a negatively skewed distribution, the left tail (the smaller values) is longer than the right tail (the larger values). This means that most of the data points are concentrated on the higher end of the scale, and a few smaller values stretch the distribution to the left.

Characteristics:

- **Mean < Median < Mode**.
- The mean is pulled towards the left because of the few extreme lower values.

Example: Age at retirement, where most people retire at the typical age, but a few may retire earlier, causing the distribution to be negatively skewed.

6.2.2 Symmetrical, Positively Skewed, and Negatively Skewed Curves

Symmetrical Curve: A symmetrical distribution has a bell-shaped curve where the left and right halves are identical.

Example: Normal distribution (Gaussian curve), where the mean, median, and mode all coincide at the center.

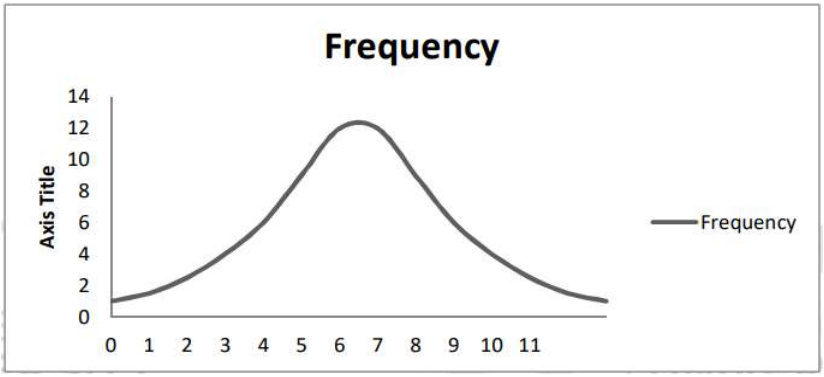


Figure: Symmetrical Curve

Source: https://www.igntu.ac.in/eContent/IGNTU-eContent-467281593500-B.Com-4-Prof.ShailendraSinghBhadouriaDean&-BUSINESSSTATISTICS-All.pdf

Positively Skewed Curve (Right Skewed): In a positively skewed distribution, the right tail is longer. The majority of data points are concentrated on the lower end of the scale.

Example: Most people earn average or lower incomes, but a few high earners pull the mean to the right.

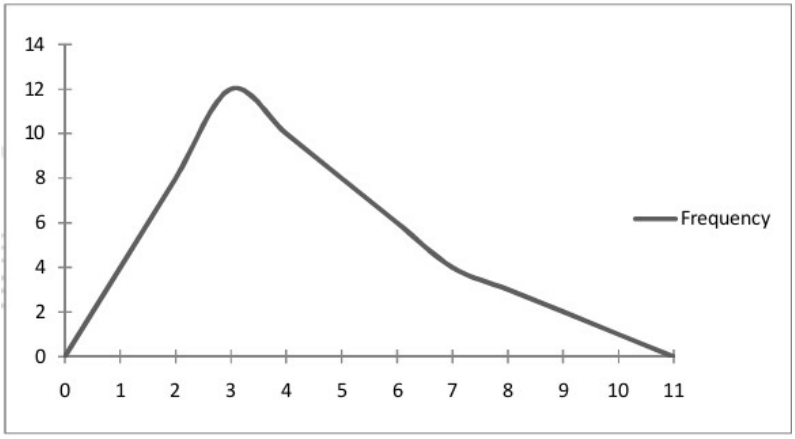


Figure: Positively Skewed Curve

Source: https://www.igntu.ac.in/eContent/IGNTU-eContent-467281593500-B.Com-4-Prof.ShailendraSinghBhadouriaDean&-BUSINESSSTATISTICS-All.pdf

Negatively Skewed Curve (Left Skewed): In a negatively skewed distribution, the left tail is longer. The majority of data points are concentrated on the higher end of the scale.

Example: In retirement age data, most people retire at a standard age, but a few early retirees cause the distribution to be skewed left.

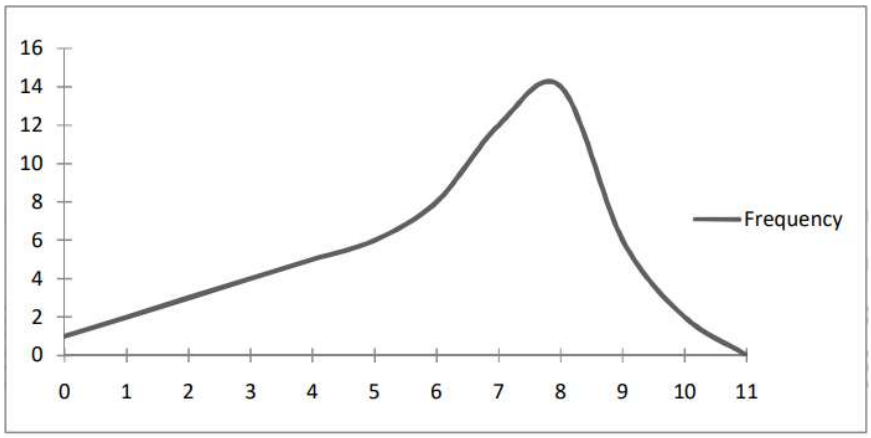


Figure: Negatively Skewed Curve

Source: https://www.igntu.ac.in/eContent/IGNTU-eContent-467281593500-B.Com-4-Prof.ShailendraSinghBhadouriaDean&-BUSINESSSTATISTICS-All.pdf

6.2.3 Various Measures of Skewness

Skewness quantifies the asymmetry of a data distribution. The measures of skewness can be divided into absolute measures and relative measures, both of which help assess how lopsided a distribution is.

Absolute Measures of Skewness

Absolute measures of skewness directly measure the difference between the mean, median, and mode. These measures are more intuitive, as they reflect the direction and degree of skewness but do not normalize the scale of variation in the data.

Skewness (Sk) = Mean – Median.

This measure calculates the difference between the mean and median. If the mean is greater than the median, the distribution is positively skewed (right-skewed). If the mean is less than the median, the distribution is negatively skewed (left-skewed).

Example: Suppose a company tracks employee salaries in a small firm: BDT 30,000, BDT40,000, BDT45,000, BDT50,000, and BDT150,000.

Mean $= \frac{30,000+4\ ,000+45,000+50,000+150,000}{5}$

$= 63,000$

Median = 45,000

Skewness = 63,000 − 45,000 = 18,000

Interpretation: Since the mean is greater than the median, the distribution is positively skewed (right-skewed), indicating the presence of a few high salaries pulling the mean to the right.

Skewness (Sk) = Mean - Mode

This measure calculates the difference between the mean and mode. If the mean is greater than the mode, the distribution is positively skewed, and if the mean is less than the mode, the distribution is negatively skewed.

Example: Consider the dataset of customer purchase amounts in a store: 5, 5, 7, 5, 8, 10, and 30.

Mean = $\frac{5+5+7+5+8+10+30}{7}$ = 10.

Mode = 5 (the most frequent value).

Skewness = 10 – 5 = 5

Interpretation: The mean is greater than the mode, indicating that the distribution is positively skewed, and there might be some higher purchase amounts that increase the mean.

Skewness (Sk) = $(Q_3 - Q_2) - (Q_2 - Q_1)$

This measure involves the differences between quartiles (Q_1, Q_2, and Q_3). It is useful when the distribution is not easily represented by just the mean, median, and mode. Positive skewness means the data has a longer tail on the right side, and negative skewness indicates a longer left tail.

Example: For a dataset of exam scores: 30, 40, 50, 60, 70, 80, and 90.

Q_1 (first quartile) = 40, Q_2 (median) = 60, Q_3 (third quartile) = 80.

Skewness, Sk = (80−60) − (60−40) = 20−20 = 0

Interpretation: Since the value is zero, the distribution is symmetric.

Relative Measures of Skewness

Relative measures of skewness normalize the skewness value, allowing for a better comparison between distributions, even if they have different units or scales. These measures are independent of the data's scale, which makes them particularly useful in comparing distributions from different datasets.

Karl Pearson's Coefficient of Skewness: This measure normalizes the difference between the mean and median by the standard deviation. A positive skew indicates that the mean is greater than the median, and a negative skew indicates the opposite, and the value of this coefficient would be zero if the distribution is symmetric.

$Sk_P = \frac{3\ (\text{Mean}-\text{Media}\ )}{\text{Standard Deviation}}$

Example: Consider the ages of a group of students: 18, 19, 20, 21, and 22.

Mean = 20, Median = 20, Standard Deviation = 1.41.

Skewness = $\frac{3(20-20)}{1.41}$

= 0

Interpretation: The skewness is zero, indicating that the distribution is symmetric.

Bowley's Coefficient of Skewness: This method uses quartiles (Q_1, Q_2, and Q_3) to calculate skewness. A positive value indicates a positively skewed distribution and a negative value indicates a negatively skewed distribution.

$Sk_B = \frac{Q_3-2Q_2+Q_1}{Q_3-Q_1}$

Example: Consider a dataset of house prices: BDT200,000, BDT250,000, BDT300,000, BDT350,000, BDT400,000.

Q_1 = 250,000, Q_2 (Median) = 300,000, Q_3 = 350,000.

$$Sk_B = \frac{350{,}000-(2\times300{,}000)+25\ ,000}{350{,}000-250{,}000}$$

$$= 0$$

Interpretation: The skewness is zero, indicating symmetry in the house prices.

Kelly's Coefficient of Skewness: Kelly's skewness is based on percentiles or deciles and helps assess the distribution's asymmetry by comparing the spread between the 90th, 50th, and 10th percentiles (or similar deciles).

$$Sk_K = \frac{(P_{90}-P_{50})-(P_{50}-P_{10})}{P_{90}-P_{10}} = \frac{(P_{90}-2P_{50}+P_{10})}{P_{90}-P_{10}}$$

Example:

A dataset of monthly expenses of 100 households: 10th percentile (P_{10}) = BDT 500, 50th percentile (P_{50}) = BDT 1,000, 90th percentile (P_{90}) = BDT 2,000.

$$Sk_K = \frac{2{,}000-2\times1{,}000+500}{2{,}000-500}$$

$$= 0.33$$

Interpretation: The positive skewness value indicates that the distribution is positively skewed, with a few high-expense households pulling the tail to the right.

Measuring Skewness for Grouped Data

Illustration

Consider the following frequency distribution of test scores for a class of students:

Score	Frequency (F)
0-10	5
10-20	10
20-30	15
30-40	8
40-50	2

Determine Skewness.

Solution

Score	Midpoint (x)	Frequency (f)	Cf	fx	$(x-\overline{x})^2$	$f(x-\overline{x})^2$
0-10	5	5	5	25	324	1620
10-20	15	10	15	150	64	640
20-30	25	15	30	375	4	60
30-40	35	8	38	280	144	1152
44-50	45	2	40	90	484	968
		N= 40		$\sum fx$=920		$\sum f(x-\overline{x})^2$=4440

Karl Pearson's Coefficient of Skewness

Mean, $\overline{x} = \frac{920}{40}$

$= 23$

The median class is the $\frac{N}{2} = \frac{40}{2} = 20^{th}$ student, which falls in the class 20-30.

Median $= 20 + (\frac{20-15}{15}) \times 10$

$= 23.33$

Variance $= \frac{4440}{40}$

$= 111$

Standard Deviation $= \sqrt{111}$

$= 10.54$

$Sk = \frac{3\,(23-23.33)}{10.54}$

$= -0.094$

Interpretation: The skewness is negative, indicating a slightly left-skewed distribution (the left tail is longer than the right tail).

Bowley's Coefficient of Skewness

$Q_1 = 10 + (\frac{\frac{40}{4}-5}{10}) \times 10$

$= 15$

Q_2 (Median) $= 23.33$

$Q_3 = 20 + (\frac{\frac{3\times40}{4}-15}{15}) \times 10$

$= 30$

$Sk_B = \frac{30-2\,(23.33)+15}{30-15}$

$= -0.11$

Interpretation: The skewness is negative, indicating that the distribution is left-skewed (the left tail is longer than the right tail).

Kelly's Coefficient of Skewness

$P_{10} = 0 + (\frac{\frac{10\times40}{100}-0}{5}) \times 10$

$= 8$

$P_{50} = 20 + (\frac{\frac{50\times40}{100}-15}{15}) \times 10$

$= 23.33$

$P_{90} = 30 + (\frac{\frac{90\times40}{100}-30}{8}) \times 10$

$= 37.5$

$Sk_K = \frac{(P_{90}-2P_{50}+P_{10})}{P_{90}-P_{10}} = \frac{37.5-2\times23.33+8}{37.5-8} = \frac{-1.16}{29.5}$

$= -0.039$

Interpretation: The negative skewness indicates a left-skewed distribution, with the lower values pulling the distribution to the left.

6.3 Moments

In statistics, moments are quantitative measures used to describe the shape of a data distribution. They provide important information about the central tendency, dispersion, symmetry, and tail behavior of a dataset. Moments are calculated about a certain point, typically the mean or origin. The Greek letter μ (mu) is used to denote moments.

Moments about Mean

For Ungrouped Data

$\mu_1 = \frac{\sum(x-\bar{x})}{N}$

$\mu_2 = \frac{\sum(x-\bar{x})^2}{N}$

$\mu_3 = \frac{\sum(x-\bar{x})^3}{N}$

$\mu_4 = \frac{\sum(x-\bar{x})^4}{N}$

For grouped Data

$\mu_1 = \frac{\sum f(x-\bar{x})}{N}$

$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N}$

$\mu_3 = \frac{\sum f(x-\bar{x})^3}{N}$

$\mu_4 = \frac{\sum f(x-\bar{x})^4}{N}$

Moments about Arbitrary Point

For Ungrouped Data

$\mu'_1 = \frac{\sum(x-A)}{N}$

$\mu'_2 = \frac{\sum(x-A)^2}{N}$

$\mu'_3 = \frac{\sum(x-A)^3}{N}$

$\mu'_4 = \frac{\sum(x-A)^4}{N}$

For grouped Data

$\mu'_1 = \frac{\sum f(x-A)}{N}$

$\mu'_2 = \frac{\sum f(x-A)^2}{N}$

$\mu'_3 = \frac{\sum f(x-A)^3}{N}$

$\mu'_4 = \frac{\sum f(x-A)^4}{N}$

With the help of following relationships, moments about arbitrary points can be converted into moments about the mean

$\mu_1 = 0$

$\mu_2 = \mu'_2 - (\mu'_1)^2$

$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3$

$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'^2_1\mu'_2 - 3(\mu'_1)^4$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

β_1 is used as a measure of skewness. However, as it uses the square of the third central moment, it is always non-negative. It indicates the degree of skewness, but not the direction (left or right). Therefore, the coefficient of skewness based on moments is defined as:

$\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3}$

where: $\mu_3 = \frac{1}{n}\Sigma(x - \bar{x})^3$ is the third central moment.

$\sigma = \sqrt{\mu_2}$ is the standard deviation.

Thus, γ_1 gives both direction and degree of skewness, while β_1 only gives the degree.

Practical Problems:

Ungrouped data: Consider the ages of a group of students: 18, 19, 20, 21, 22. Calculate the first four moments.

Mean $= \frac{18+19+20+21}{5}$

$= 20$

$\mu_1 = \frac{(18-20)+(19-20)+(20-2\ )+(21-20)+(22-20)}{5}$

$= 0$

x	$(x - \bar{x})^2$
18	4
19	1
20	0
21	1
22	4
Total	10

$\mu_2 = \frac{10}{5}$

$= 2$

x	$(x - \bar{x})^3$
18	8
19	1
20	0
21	1
22	8
Total	18

$\mu_3 = \frac{18}{5}$

$= 3.6$

x	$(x - \bar{x})^4$
18	16
19	1
20	0
21	1
22	16
Total	34

$\mu_4 = \frac{34}{5}$

$= 6.8$

6.4 Kurtosis

Kurtosis is a statistical measure that describes the tailedness or sharpness of the peak of a distribution. It provides insight into the extreme values (outliers) present in the data. While measures like the mean and variance describe the central tendency and spread, kurtosis focuses specifically on the shape of the distribution, particularly its tails.

In simpler terms, kurtosis tells us whether the data has:

- Heavy tails (many outliers),
- Light tails (fewer outliers),
- Or if it behaves similarly to a normal distribution.

6.4.1 Types of Kurtosis:

Leptokurtic: Distributions with high kurtosis are called leptokurtic. These distributions have heavy tails and sharp peaks. They tend to have more extreme values (outliers). It means Kurtosis > 3 (or positive excess kurtosis, > 0).

Example: Stock market returns (with extreme positive or negative fluctuations).

Mesokurtic: Distributions with normal kurtosis are called mesokurtic. They have a bell-shaped curve similar to a normal distribution, with moderate tails. It means Kurtosis = 3 (or zero excess kurtosis).

Example: The standard normal distribution.

Platykurtic: Distributions with low kurtosis are called platykurtic. These distributions have light tails and a flatter peak compared to a normal distribution. It means Kurtosis < 3 (or negative excess kurtosis, <0).

Example: Uniform distribution, where all values are equally likely, and extreme values are rare.

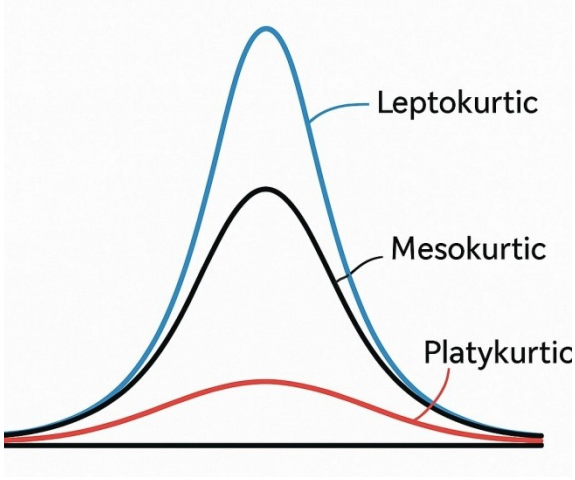


Figure: Types of kurtosis

6.4.2 Why Kurtosis Matters:

Kurtosis is important because it helps you understand the risk of extreme deviations from the mean. High kurtosis indicates that the data may have more extreme values than a normal distribution, while low kurtosis suggests fewer outliers. This is particularly useful in fields like finance, quality control, and risk management where extreme events can have significant impacts.

Formula for Kurtosis: $\beta_2 = \frac{\mu_4}{\mu_2^2}$

Where, μ_4 = fourth central moment

μ_2= second central moment (variance)

Example: The first central moments are 0, 16, -36, and 120. Determine kurtosis.

Solution: Given $\mu_1 = 0$, $\mu_2 = 16$, $\mu_3 = -36$, and $\mu_4 = 120$

Kurtosis = $\frac{120}{16^2}$

$= 0.469$

Since the value is below 3, the distribution is platykurtic.

Grouped Data Illustration:

Consider the following frequency distribution of test scores for a class of students:

Score	Frequency (*f*)
0-10	5
10-20	10
20-30	15
30-40	8
40-50	2

Determine the first four moments, skewness, and kurtosis.

Solution

Score	Midpoint (x)	Frequency (f)	cf	fx	$(x-\bar{x})$	$f(x-\bar{x})$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
0-10	5	5	5	25	-18	-90	1620	-29160	524880
10-20	15	10	15	150	-8	-80	640	-5120	40960
20-30	25	15	30	375	2	30	60	120	240
30-40	35	8	38	280	12	96	1152	13824	165888
40-50	45	2	40	90	22	44	968	21296	468512
		N= 40		920		$\sum f(x-\bar{x})=0$	$\sum f(x-\bar{x})^2=4440$	$\sum f(x-\bar{x})^3=960$	$\sum f(x-\bar{x})^4=1200480$

Mean, $\bar{x} = \frac{920}{40}$

$= 23$

$\mu_1 = \frac{0}{40}$

$= 0$

$\mu_2 = \frac{4440}{40}$

$= 111$

$\mu_3 = \frac{960}{40}$

$= 24$

$\mu_4 = \frac{1200480}{40}$

$= 30012$

Skewness, $\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{24}{111^{3/2}} = 0.021$

The skewness value is very close to zero, which indicates that the distribution is nearly symmetric. A small positive skew suggests a very slight tendency for the data to have a longer right tail, but this asymmetry is quite minor.

Kurtosis, $\beta_2 = \frac{30012}{111^2} = 2.44$

Since the kurtosis value 2.44 is less than 3, the distribution is platykurtic, meaning it has lighter tails and fewer extreme outliers than a normal distribution.

Review Questions

1. Define skewness. Explain the types of skewness.
2. What is skewness? What does it tell us about a distribution?
3. What are the absolute and relative measures of skewness?
4. Write the formula for Pearson's coefficient of skewness and interpret its values.
5. What does a positive value of skewness indicate?
6. How can you calculate skewness using the third central moment?
7. What is a statistical moment? Briefly describe the first four moments.
8. How is the mean related to the first moment?
9. Define central moments. How do they differ from raw moments?
10. Define kurtosis in statistics. Discuss the types of kurtosis.
11. What is the formula for kurtosis using moments? Discuss the importance of Kurtosis.
12. Differentiate between mesokurtic, leptokurtic, and platykurtic distributions.
13. What does a kurtosis value less than 3 indicate about the shape of a distribution?
14. Consider the following grouped data representing the monthly salaries of employees in a company:

Salary Range (x)	Frequency (f)
1000 - 2000	4
2001 - 3000	7
3001 - 4000	10
4001 - 5000	6

Calculate Karl Pearson's coefficient of skewness and Kelly's coefficient of skewness.

15. Consider the following grouped data representing the monthly expenses of households:

Expense (x)	Frequency (f)
10000 - 20000	40
20001 - 30000	70
30001 - 40000	100
40001 - 50000	60

Calculate the first four moments and kurtosis.

16. Given a distribution with $\mu_2=25$ and $\mu_4=2000$, calculate the kurtosis and identify the type of distribution.
17. The mean of a dataset is 60, the median is 55, and the standard deviation is 10. Find the Pearson's coefficient of skewness.
18. A frequency distribution has $\mu_2=20$, $\mu_3=0$, and $\mu_4=1800$.
    a. Is the distribution symmetric?
    b. What type of kurtosis does it have?

References

Gupta, S. P., & Gupta, M. P. (2020). *Business statistics*. Sultan Chand & Sons.

Fisher, R. A. (1934). Statistical Methods for Research Workers. *Oliver and Boyd*.

Pearson, K. (1900). The Grammar of Science. *Walter Scott Publishing Co.*

Bhadouria, S. S. (Dean). (2023). *Business statistics* (B.Com 4). Indira Gandhi National Tribal University. Retrieved from https://www.igntu.ac.in/eContent/IGNTU-eContent-467281593500-B.Com-4-Prof.ShailendraSinghBhadouriaDean&-BUSINESSSTATISTICS-All.pdf

CORRELATION ANALYSIS

7

Unit Highlights

- Introduction
- Significance of the Study of Correlation
- Correlation and Causation
- Types of Correlation
- Methods of Correlation

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 7: Correlation Analysis

Learning Objectives

By the end of this Unit, you should be able to:

1. Understand the concept of correlation and explain its significance in analyzing relationships between variables.
2. Differentiate between correlation and causation, and evaluate their implications in data analysis.
3. Identify and classify types of correlation based on direction, strength, and linearity.
4. Apply various methods of correlation analysis, including the scatter diagram method, Karl Pearson's coefficient of correlation, and Spearman's rank correlation.
5. Interpret the results of correlation analysis, including the coefficient of correlation, and discuss its properties and practical applications.

7.1 Introduction

Correlation analysis is a powerful statistical tool for examining relationships between variables, providing insight into how changes in one factor might be associated with changes in another (Granger, 1969). In the real world, few phenomena exist in isolation; most involve interconnected variables that move together in some way. For instance, as family income rises, people might spend more on luxury items, or as the price of a commodity falls, demand for it might increase. Correlation analysis helps us quantify these relationships, enabling us to understand the strength and direction of the associations we observe.

The text defines correlation as a statistical measure of covariation, meaning it evaluates the degree to which two or more variables move together. This relationship is quantified by the correlation coefficient (symbolized by r), which ranges from -1 to 1. A positive r-value suggests that as one variable increases, the other does too, while a negative r-value indicates an inverse relationship (Pearson, 1895). For example, a positive correlation might be observed between family income and luxury spending, while a negative correlation might be found between the price of a commodity and its demand.

To effectively analyze correlations, a three-step approach is often recommended. First, it's essential to determine whether a relationship exists between the variables. This step involves calculating the correlation coefficient and visually inspecting the data to see if patterns emerge. Second, we must test the significance of the relationship to ensure it's not due to random chance. Statistical tests allow us to assess if the observed correlation is strong enough to be considered meaningful. Finally, while correlation does not imply causation, exploring the possibility of a causal link can be beneficial for a deeper understanding. However, it is crucial to remember that a significant correlation alone is not proof of causation.

A common misconception about correlation is that a strong relationship between two variables implies one causes the other. This is not necessarily the case. For instance, an analysis might reveal a strong correlation between smoking and lung cancer. While this relationship is significant, it does not inherently prove that smoking causes cancer. There could be additional factors, or confounders, influencing both smoking behavior and cancer risk. This is why researchers emphasize caution when interpreting correlations, especially in fields where multiple variables interact in complex ways.

Reliable correlation analysis depends on having paired observations across variables. Each data pair represents the values of two variables measured under the same conditions, whether based on time, location, or other controlled settings. For example, if we wanted to study the relationship between study hours and test scores among students, we would need each student's study time paired with their corresponding test result. Without paired observations, correlation results could be misleading or invalid.

7.2 Significance of the Study of Correlation

The study of correlation is significant because it provides insights into the relationships between variables, helping researchers, analysts, and decision-makers understand and predict patterns in data. Here are some key reasons why correlation analysis is essential:

1. Identifying Relationships Between Variables

Correlation helps to reveal associations between two or more variables, which can guide understanding and interpretation. For example, discovering a correlation between advertising expenditure and sales can help a business understand the potential impact of marketing efforts on revenue.

2. Predictive Value

Once a significant correlation is established, it can be used to make predictions. In fields like finance, if a stock's performance is correlated with certain economic indicators, analysts can use these correlations to predict future price movements. Similarly, in healthcare, if certain behaviors are correlated with health outcomes, this information can be used to predict and improve patient health.

3. Informing Decision-Making

By identifying correlations, organizations and policymakers can make informed decisions. For instance, if studies show a positive correlation between education levels and employment rates, a government might invest more in education to reduce unemployment. Correlation analysis provides a basis for decisions that aim to improve outcomes based on observed relationships.

4. Understanding Complex Phenomena

Many real-world phenomena are influenced by multiple variables interacting in complex ways. Correlation analysis can help break down these complexities by identifying specific pairs of variables that are associated, providing a clearer picture of how different factors interrelate. For example, in environmental science, correlations between pollution levels and respiratory health can contribute to understanding the broader impacts of pollution on communities.

5. Highlighting Potential Causal Relationships

Although correlation does not imply causation, finding a strong correlation can point researchers in the direction of a possible causal relationship worth exploring. For example, a strong correlation between physical activity and lower rates of certain diseases suggests an area for further research into potential causal effects. This preliminary insight often forms the basis for more rigorous experiments and studies.

6. Building Statistical Models

Correlation is often the first step in developing more complex statistical models, such as regression analysis. By understanding correlations between variables, analysts can build

models that predict or explain outcomes, which are widely used in fields like economics, psychology, and data science. For instance, a correlation between education and income can lead to a regression model that more accurately predicts income levels based on educational attainment.

7. Guiding Further Research

Correlation studies often serve as a foundation for further research. When a significant correlation is found, researchers may investigate the underlying mechanisms or expand the study to examine other variables that could influence the relationship. For instance, a correlation between screen time and mental health issues in children might lead to studies on how different types of screen content affect well-being.

The study of correlation is crucial across many fields because it offers a practical way to understand relationships, make predictions, and inform decisions. While it does not establish causation, correlation analysis lays the groundwork for further research and provides valuable insights that help address real-world problems.

7.3 Correlation and Causation

Understanding the difference between correlation and causation is crucial for interpreting relationships between variables accurately. Imagine a scenario where an economist observes that as ice cream sales increase, so do incidents of sunburn. At first glance, it might seem as though eating ice cream somehow leads to sunburn, especially since the numbers appear to rise and fall together. However, this correlation does not mean that ice cream causes sunburn. Instead, both are influenced by a third factor – warmer weather. As temperatures rise, people are more likely to buy ice cream to cool down and to spend time outside, increasing their risk of sunburn. Here, warmer weather acts as a "confounding variable" that influences both ice cream sales and sunburn rates, leading to a correlation without a direct causal link.

In other situations, mutual dependence between variables can create correlations that might seem like causation but reflect an interplay. Consider supply and demand in economics. When the price of a commodity rises, demand often decreases because fewer people are willing to buy at a higher cost. Conversely, when demand falls, prices might decrease as sellers try to attract more buyers. Both factors – supply and demand – are interdependent, constantly adjusting based on the other. While they appear correlated, it's difficult to pinpoint which factor is causing the other; instead, they mutually influence each other in a dynamic balance. This mutual dependence is common in economic and business data, where variables often impact each other in a cycle rather than in a one-way cause-and-effect relationship.

Another intriguing example is when two variables correlate purely by chance, leading to what is known as a "spurious correlation." For instance, if a study finds a correlation between shoe size and intelligence within a certain population, it does not imply that larger shoe sizes lead to higher intelligence. This correlation might be due to random variation within a small sample, or it might be a complete coincidence. Such spurious correlations highlight the danger of assuming that any observed association has meaningful implications. In reality, these connections can be random, occurring only because of chance patterns within the data rather than any real-world relationship.

These examples emphasize why correlation alone is not enough to prove causation. True causation requires rigorous evidence, often through controlled experiments that isolate variables to show a direct cause-and-effect relationship. For instance, to demonstrate that

smoking causes lung cancer, researchers conducted studies over many years, accounting for lifestyle and genetic factors, and consistently observed higher cancer rates among smokers. Unlike simple correlation, establishing causation involves proving that one factor directly influences another, ruling out confounding variables and random chance.

In everyday decision-making, distinguishing between correlation and causation helps prevent misinterpretation of data. If policymakers mistakenly assume that ice cream sales cause sunburn, they might impose unnecessary restrictions on ice cream vendors rather than focusing on sunscreen use and sun safety. By carefully analyzing whether relationships between variables are correlational or causal, researchers, economists, and policymakers can make better-informed decisions, accurately address issues, and avoid the pitfalls of false assumptions.

7.4 Types of Correlation

1. Types of Correlation Based on Direction

The direction of correlation indicates whether the variables move in the same or opposite directions.

a. Positive Correlation: In a positive correlation, as one variable increases, the other variable also increases, and vice versa. For example, there is often a positive correlation between the number of hours studied and test scores – as study hours increase, test scores also tend to improve.

b. Negative Correlation: In a negative correlation, as one variable increases, the other variable decreases. For example, there may be a negative correlation between the price of a commodity and its demand – as the price goes up, demand typically goes down.

c. No Correlation (Zero Correlation): When there is no consistent relationship between two variables, they are said to have no correlation or zero correlation. For instance, there may be no correlation between a person's shoe size and their intelligence level.

2. Types of Correlation Based on Strength

The strength of correlation refers to how closely the variables are related. Correlation coefficients, that range from -1 to +1, are used to quantify this strength.

a. Perfect Correlation: A perfect correlation occurs when the correlation coefficient is exactly +1 or -1, meaning the variables have a completely consistent and predictable relationship. In a perfect positive correlation (+1), variables move exactly in sync; in a perfect negative correlation (-1), they move exactly opposite to each other.

b. Strong Correlation: A strong correlation means that the variables are closely related, but the relationship is not perfect. Correlation coefficients close to +1 or -1, such as 0.8 or -0.8, indicate a strong positive or strong negative correlation, respectively.

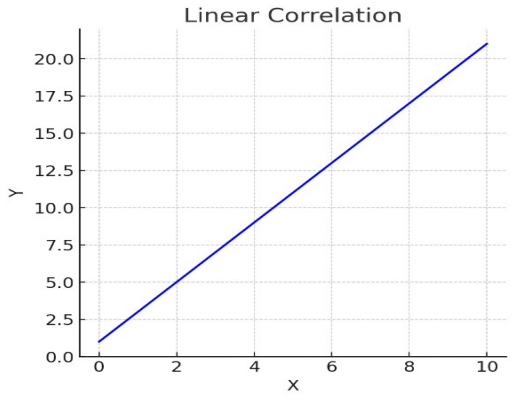
c. Moderate Correlation: A moderate correlation means the variables have some association, but it's not as close as a strong correlation. Correlation coefficients around 0.5 or -0.5 represent moderate correlations.

d. Weak Correlation: A weak correlation suggests that the relationship between the variables is minimal or inconsistent. Correlation coefficients closer to 0, such as 0.2 or -0.2, indicate weak correlations.

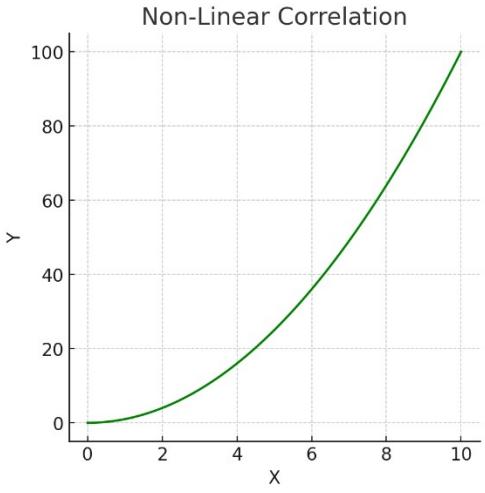
3. Types of Correlation Based on Linearity

The linearity of correlation refers to the shape of the relationship when plotted on a graph.

a. Linear Correlation: In a linear correlation, the relationship between the two variables forms a straight line when plotted on a scatter plot. This means that the change in one variable is proportionally consistent with the change in the other variable. For example, height and weight often have a linear positive correlation in adults.



b. Non-Linear (Curvilinear) Correlation: In a non-linear or curvilinear correlation, the relationship between the variables is not consistent throughout. This means that changes in one variable may not correspond to a proportional change in the other variable, resulting in a curve rather than a straight line. For example, the relationship between age and physical strength can be curvilinear, where strength increases in youth, peaks in adulthood, and declines in old age.



Type of Correlation	Description
Positive Correlation	Both variables increase or decrease together.
Negative Correlation	One variable increases while the other decreases.
No Correlation	No consistent relationship between variables.
Perfect Correlation	Variables have an exact, consistent relationship (correlation coefficient of ±1).
Strong Correlation	Variables are closely related, with a high correlation coefficient.
Moderate Correlation	Variables are somewhat related, with a moderate correlation coefficient.
Weak Correlation	Variables have a minimal, inconsistent relationship.
Linear Correlation	The relationship forms a straight line on a scatter plot.
Non-Linear Correlation	The relationship forms a curve on a scatter plot.

7.5 Methods of Correlation

There are several methods for studying correlation, each providing insights into the relationship between two or more variables:

-Scatter Diagram Method
-Karl Pearson's Coefficient of Correlation
-Spearman's Rank Correlation coefficient

7.5.1 Scatter Diagram Method

A scatter diagram provides a vital, approachable tool for examining relationships between two quantitative variables. It is a graphic portrayal of how two characteristics or metrics interact, where each dot or point on the graph represents an individual data observation, positioned by its values for each variable. With the first variable on the x-axis and the second on the y-axis, a scatter diagram immediately reveals patterns, trends, and potential relationships that might otherwise remain hidden in a simple list of numbers.

One of the scatter diagram's most insightful aspects is its ability to distinguish between types of correlation: positive, negative, curvilinear, and none. A positive linear correlation reveals itself in an upward trend, with points roughly aligning along an imaginary line ascending from the lower left to the upper right. This indicates that as one variable increases, so does the other. For example, consider a study of education level and income; individuals with higher education levels often correlate with higher incomes, and this relationship would typically manifest as a positive linear trend on a scatter plot.

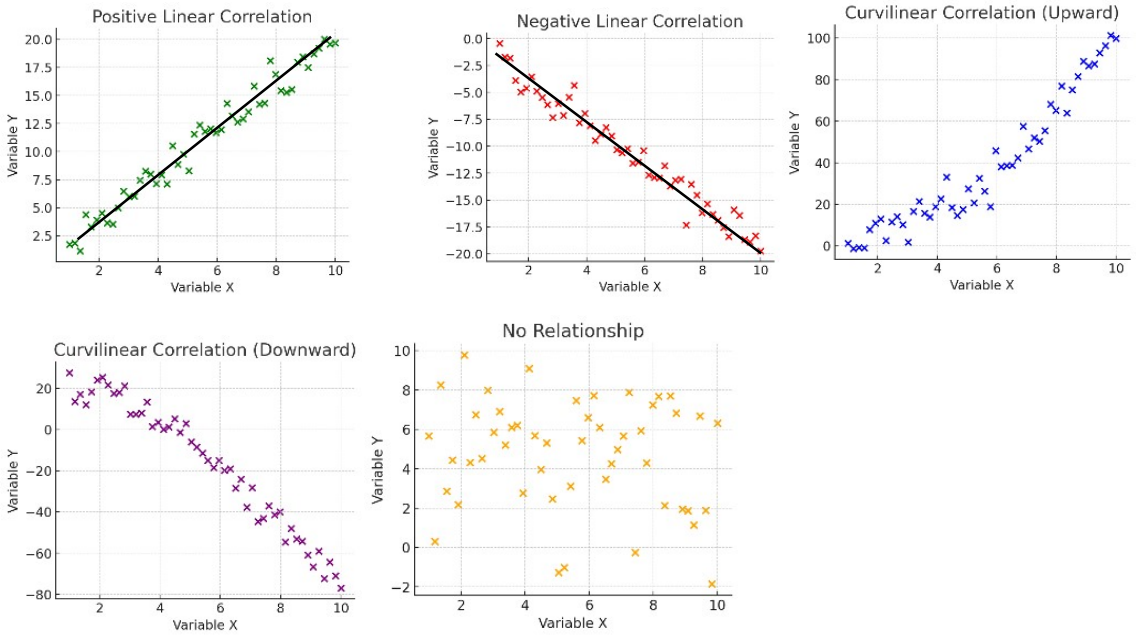


Figure: Scatter diagram

In contrast, a negative linear correlation shows a descending pattern from the upper left to the lower right, indicating that as one variable increases, the other decreases. For instance, a negative correlation might appear in data comparing the hours spent on leisure and stress levels, suggesting that more leisure time aligns with reduced stress.

Scatter diagrams can also reveal curvilinear correlations, where the relationship isn't simply positive or negative but instead follows a curve. For example, a study of productivity versus hours worked might initially show productivity increasing with hours, but after a point, productivity begins to decline as fatigue sets in. This pattern would appear as a curve rather than a straight line.

Equally important is the potential to find no correlation, where points scatter randomly across the graph, revealing no discernible relationship between the variables. This lack of correlation is often just as informative, demonstrating that some variables operate independently without influencing each other.

However, an important caution accompanies scatter diagrams: correlation is not causation. Just because two variables trend together does not imply one causes the other. For example, a scatter diagram might show an upward trend between ice cream sales and sunscreen sales, yet this does not mean that ice cream causes sunscreen purchases. Instead, a third factor, such as warm weather, is likely driving both trends.

7.5.2 Karl Pearson's Coefficient of Correlation

Karl Pearson's coefficient of correlation, often symbolized as *r*, is a widely used statistical measure that quantifies the strength and direction of the linear relationship between two variables. This coefficient is central in correlation analysis, developed by the British mathematician Karl Pearson. It forms the foundation of many statistical methods, especially in economics, psychology, and natural sciences.

The formula for Pearson's correlation coefficient between two variables X and Y is:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 . \sum(Y_i - \overline{Y})^2}}$$

where:

- X_i and Y_i are individual observations of the variables *X* and *Y*.
- $\overline{X}$ and $\overline{Y}$ are the means of *X* and *Y*.
- The numerator represents the covariance between *X* and *Y*, and the denominator normalizes this covariance by the product of the standard deviations of *X* and *Y*.

The above formula can be written as:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}}$$

where:

$$x = (X_i - \overline{X})$$
$$y = (Y_i - \overline{Y})$$

Interpretation of the coefficient of Correlation

The result, *r*, ranges from -1 to +1:

r = +1: Perfect positive linear relationship (as one variable increases, the other increases proportionally).
r = −1: Perfect negative linear relationship (as one variable increases, the other decreases proportionally).
r = 0: No linear relationship between the variables.

The closer r is to ±1, the stronger the linear relationship. Values close to 0 suggest weak or no linear correlation. A positive r indicates a positive correlation, while a negative r signals a negative correlation.

Pearson's coefficient requires that:

1. The data is continuous and approximately normally distributed.
2. The relationship between the variables is linear.
3. Outliers are minimal, as they can skew the correlation value significantly.

Properties of the Coefficient of Correlation

The coefficient of correlation, particularly Pearson's r, has several properties that make it a valuable and insightful tool in statistical analysis. These properties help interpret the coefficient's value and understand its implications in data analysis. Here's a summary of its core properties:

1. The value of the coefficient of correlation r is always between -1 and +1: $-1 \leq r \leq 1$
   An r of +1 indicates a perfect positive linear relationship, while -1 indicates a perfect negative linear relationship. An r of 0 means there is no linear relationship between the variables.
2. The correlation coefficient is symmetric, meaning that the correlation between X and Y is the same as the correlation between Y and X. $r_{XY}=r_{YX}$
3. r is a dimensionless quantity, meaning it has no units. It provides a standardized measure of the relationship, allowing comparison across different datasets regardless of the scale of measurement of the variables.
4. Pearson's correlation coefficient measures only the strength of a linear relationship. If the relationship between two variables is non-linear, r may not accurately reflect the strength of association, even if the variables are strongly related in a non-linear way.
5. Multiplying either or both variables by a positive constant does not change the correlation coefficient.
   Adding or subtracting a constant from either variable also does not affect r. Thus, the correlation coefficient is invariant to changes in the scale or origin of the data.
6. A high or low correlation does not imply causation between the two variables. Correlation only measures the degree to which two variables move together, not whether one variable influences the other.
7. Pearson's r is sensitive to outliers, as extreme values can disproportionately affect the calculation of the correlation. An outlier may create a misleadingly high or low correlation, even if the overall data pattern is different.
8. Pearson's r is closely related to covariance; it is essentially a scaled version of the covariance. Specifically:
   $$r = \frac{\mathrm{Cov(X,Y)}}{\sigma_X \sigma_Y}$$
   This property helps in interpreting r as a standardized form of covariance, adjusted to fit within the range of -1 to +1.
9. If two variables are independent, then their correlation coefficient is 0. However, a correlation coefficient of 0 does not necessarily imply independence; it merely indicates no linear relationship. Non-linear relationships can still exist even when $r = 0$.

Illustration: Using the following data:

a. Calculate the coefficient of correlation

b. Estimate the percentage of the group with lung cancer in a country where 15 per cent of the group smoke heavily.

Country	% of group smoking heavily	% of the group with lung cancer
A	10	5
B	20	15
C	20	20
D	30	25
E	30	20

Solution: (a)

Test No.	Country	% of group smoking heavily (X)	% of the group with lung cancer (Y)	x=(X-$\bar{X}$) x=(X-22)	x^2	y=(Y-$\bar{Y}$) y=(Y-17)	y^2	xy
1	A	10	5	-12	144	-12	144	144
2	B	20	15	-2	4	-2	4	4
3	C	20	20	-2	4	3	9	-6
4	D	30	25	8	64	8	64	64
5	E	30	20	8	64	3	9	24
Total	N=5	$\sum X$=110	$\sum Y$=85	$\sum x = 0$	$\sum x^2$=280	$\sum y$=0	$\sum x^2$= 230	$\sum xy$= 230

We know that, $\bar{X} = \frac{\sum X}{N} = \frac{110}{5} = 22, \bar{Y} = \frac{\sum Y}{N} = \frac{85}{5} = 17$

$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = \frac{230}{\sqrt{280 \times 230}} = 0.91$

Since r = 0.91, hence, there is a high degree of positive correlation between smoking and lung cancer. Moreover, when the percentage (%) of smoking increases, then also the percentage (%) of lung cancer increases to 91% (r = 0.91).

(b) The required percentage of the group with lung cancer in a country where 15% of the group smoke heavily is (15×0.91)=13.65%.

Illustration: Calculate Karl Pearson's coefficient of correlation between expenditure on product development and revenue from the data given below.

Product Development Expenses ('000 Tk.)	45	68	55	88	76	32	95	48	81
Revenue (lakh Tk.)	50	57	60	85	70	45	89	55	78

Solution:

Product Development Expenses (X)	Revenue (Y)	x=(X-65.33)	y=(Y-65.44)	xy	x^2	y^2
45	50	-20.33	-15.44	313.9	413.3	238.5
68	57	2.67	-8.44	-22.5	7.1	71.2
55	60	-10.33	-5.44	56.2	106.7	29.6
88	85	22.67	19.56	443.6	514.7	382.5
76	70	10.67	4.56	48.7	113.8	20.8
32	45	-33.33	-20.44	681.3	1111.1	417.8
95	89	29.67	23.56	699.3	880.4	555.9
48	55	-17.33	-10.44	181.0	300.3	109.0
81	78	15.67	12.56	196.7	245.7	157.8
$\sum X$= 588	$\sum Y$=589			$\sum xy$=2598.2	$\sum x^2$ =3692.7	$\sum y^2$ =1983.1

We know that, $\bar{X} = \frac{\sum X}{N} = \frac{588}{9} = 65.33, \bar{Y} = \frac{\sum Y}{N} = \frac{589}{9} = 65.44$

$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = \frac{2598.2}{\sqrt{(3692.7).(1983.1)}} = 0.96$

The Pearson correlation coefficient, r = 0.96, indicates a strong positive correlation between product development expenses and revenue.

7.5.3 Spearman's Rank Correlation

In statistics, we often encounter cases where we want to examine the relationship between two variables. However, not all variables are numerical or quantitative. Some variables, like attributes (e.g., honesty, beauty, character), cannot be measured precisely but can be ranked in a particular order.

In situations like these, where we deal with attributes instead of measurable quantities, Karl Pearson's Correlation is not suitable because it requires numerical values. To handle this, Charles Edward Spearman, a British psychologist, developed a method in 1904 called Spearman's Rank Correlation. This method measures the correlation between the ranks of two variables, allowing us to analyze qualitative attributes (Lyerly, 1952).

Spearman's Rank Correlation (denoted as r_s) is a non-parametric measure of the strength and direction of association between two ranked variables. It assesses how well the relationship between two variables can be described using a monotonic function.

The formula is: $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$

where:

d is the difference between the ranks of each pair of values,
n is the number of pairs.

Illustration:

The scores of students in an examination in Physics and Chemistry are given below:

Student	1	2	3	4	5	6	7	8
Marks in Physics	68	50	58	62	55	48	72	60
Marks in Chemistry	60	48	52	64	56	50	70	58

Required: The Spearman rank coefficient of correlation between the scores in Physics and Chemistry.

Solution:

Student	X	Y	Rank(X)	Rank(Y)	d=Rank(X)−Rank(Y)	d^2
1	68	60	2	3	-1	1
2	50	48	7	8	-1	1
3	58	52	5	6	-1	1
4	62	64	3	2	1	1
5	55	56	6	5	1	1
6	48	50	8	7	0	0
7	72	70	1	1	0	0
8	60	58	4	4	0	0
						$\sum d^2 = 5$

$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6.5}{8(8^2-1)} = 1 - \frac{30}{504} = 1 - 0.0595 = 0.94$

A Spearman rank correlation coefficient of 0.94 indicates a strong positive association between the ranks of the two variables—in this case, scores in Physics and Chemistry. This value of 0.94 is close to 1, which means that there is a high degree of agreement in the rankings of students' scores in both subjects.

Review Questions

1. What is a correlation, and how is it used in data analysis?
2. Explain the difference between positive, negative, and zero correlation with examples.
3. What is the difference between correlation and causation? Why is it important to distinguish between the two?
4. Define Pearson's correlation coefficient. How is it interpreted?
5. What are the main characteristics of a scatter plot, and how does it help in understanding correlation?
6. Explain Spearman's rank correlation coefficient and when it should be used instead of Pearson's correlation.
7. What are the limitations of using correlation to describe the relationship between two variables?
8. What is the range of the correlation coefficient? What does each end of the range signify?
9. How can outliers affect the correlation between two variables?
10. In what situations is it better to use a rank correlation (like Spearman's) rather than Pearson's correlation coefficient?

11. If the Pearson correlation coefficient between study hours and exam scores is r= 0.85, what does this tell you about the relationship between study hours and exam scores?
12. Sketch a scatter plot that would likely represent a strong negative correlation and one that represents a weak positive correlation.
13. The scores of students in an examination in English and Science are given below:

Student No.	1	2	3	4	5	6	7	8
Marks in English	68	45	52	56	49	62	58	54
Marks in Science	64	48	55	60	52	66	59	50

Requirements:

(i) Find Correlation coefficient
(ii) Find the rank correlation coefficient and compare the two values.

14. Spearman's rank correlation coefficient between two variables is -0.9, what can be said about the relationship between these variables?

References

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438.

Lyerly, S. (1952). The average spearman rank correlation coefficient. *Psychometrika*, 17, 421-428. https://doi.org/10.1007/BF02288917.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

REGRESSION ANALYSIS

8

Unit Highlights

- Introduction
- Purposes of Regression
- Difference Between Correlation and Regression
- Types of Regression
- Simple Linear Regression

Technologies Used for Content Delivery

- BOU TUBE
- BOU LMS
- WebTV
- Web Radio
- Mobile Technology with Micro SD Card
- LP+ Office 365
- BTV Program
- Bangladesh Betar Program

Unit 8: Regression Analysis

Learning Objectives

By the end of this Unit, you should be able to:

1. Understand the purpose of regression analysis and its role in predicting relationships between variables.
2. Differentiate between correlation and regression and explain their respective uses in statistical analysis.
3. Identify and classify the types of regression, focusing on their applications.
4. Explain the concept of simple linear regression, including its objectives and assumptions.

8.1 Introduction

Regression analysis is a cornerstone of statistical methodologies, recognized for its capacity to explore and quantify relationships between variables. Regression is a pivotal tool for understanding patterns, forecasting outcomes, and supporting informed decision-making. It is widely used in economics, social sciences, engineering, and the natural sciences. At its essence, regression seeks to determine how one variable—commonly referred to as the dependent or response variable—varies with one or more independent or predictor variables.

The utility of regression is evident in its applications. For instance, an economist might analyze how interest rate fluctuations influence consumer spending, while a business strategist predicts sales based on advertising investments. Regression provides the mathematical framework to quantify these associations, enabling explanation and prediction (Berry & Feldman, 1985), (Gunst & Mason, 1980).

Formally defined, regression is a statistical technique for modeling the relationship between a dependent variable and one or more independent variables. This relationship helps estimate or predict the dependent variable's value using known values of the predictors. In its simplest form, such as linear regression, the connection is represented by a straight line that encapsulates the direction and strength of the association (Krishna et al., 2023), (Cameron & Trivedi, 1998).

By identifying these relationships, regression analysis not only deepens our understanding of variable interactions but also facilitates practical decision-making, enabling predictions and strategic planning in diverse contexts such as economic forecasting, public policy, and market analysis. Its adaptability and rigor make it a vital tool for both academic research and professional applications (Bin-jie, 2004; Lay, 2009).

Knowing the relationship between two variables, such as income (X) and consumption (Y), enables economists to predict changes in consumer spending based on income levels. Similarly, understanding the connection between the cost of production (X) and market price (Y) helps manufacturers plan pricing strategies.

Regression as a Predictive Tool

One of the most significant benefits of regression analysis is its predictive power. In addition to describing existing relationships, regression analysis is widely used to make future projections. By identifying and quantifying relationships between variables, organizations and researchers can anticipate future trends and plan accordingly. For example, in finance, regression can help forecast stock prices based on historical trends and economic indicators.

In public health, it can be used to estimate the spread of a disease based on factors like population density and vaccination rates.

8.2 Purposes of Regression

Regression analysis aims to identify, model, and quantify relationships between variables, often to predict or estimate the value of one variable based on the values of others. Here are the key purposes of regression analysis:

1. Understanding Relationships between Variables

Regression analysis is a powerful statistical tool that allows us to understand, model, and quantify relationships between variables. Imagine we're studying two factors, like advertising and sales, to see if there's a relationship between how much a company spends on advertising and how much it sells. With regression, we can go beyond just observing that there might be a connection—we can measure precisely how changes in advertising spending are likely to affect sales. This ability to quantify relationships is one of the main reasons regression analysis is so widely used in fields from economics to engineering, social sciences to medicine.

2. Prediction and Forecasting

At its core, the purpose of regression is to uncover patterns and make predictions. One of its most practical uses is in forecasting. For instance, a company might want to predict next year's sales based on factors like current advertising spend, product quality, or even economic conditions. Using historical data, regression can provide a model that shows how each factor contributes to sales, giving a realistic prediction based on past trends. This is invaluable for planning and decision-making, as it allows businesses to estimate outcomes under different scenarios—such as what might happen if they increase their advertising budget by a certain percentage.

3. Quantifying Effects and Causality

Regression quantifies how much one variable changes as another variable changes, which can help measure the strength and nature of relationships. For example, economists may use regression to quantify how much a tax rate change impacts consumer spending. While correlation doesn't imply causation, regression can provide evidence of causal relationships when combined with well-designed experiments or longitudinal data, such as understanding how educational interventions impact test scores.

4. Optimization and Decision-Making

Regression also plays a role in optimizing decisions and resources. In business, knowing which factors drive profitability can help managers make informed choices about where to invest. For instance, if regression shows that customer satisfaction has a stronger impact on revenue than advertising does, a company might choose to focus on improving customer experience. Similarly, in manufacturing, regression can help optimize production processes by analyzing how different factors like temperature or material quality affect output, leading to more efficient operations.

8.3 Difference between Correlation and Regression

Correlation measures the strength and direction of a relationship between two variables. It tells us if two variables tend to move together (either positively or negatively), but it does not imply causation. Regression, on the other hand, is about predicting or estimating the value of

one variable based on the value of another. It helps establish a relationship where one variable (independent variable) can be used to predict or explain the other (dependent variable).

Correlation only tells us whether a relationship exists and whether it is positive or negative. For example, if the correlation coefficient (r) is positive, it indicates that as one variable increases, the other also tends to increase. On the other hand, regression describes the nature of the relationship by creating an equation (regression line) that represents how one variable changes as the other changes. This relationship is often expressed in the form of a line or curve, providing a more detailed understanding.

Again, Correlation does not consider which variable is independent (cause) or dependent (effect). It simply measures the association between two variables without establishing causation. However, regression explicitly distinguishes between independent (predictor) and dependent (response) variables. In a regression analysis, the independent variable is assumed to influence or predict the dependent variable.

Correlation is typically measured by the correlation coefficient (r), which ranges from -1 to +1. Values close to +1 or -1 indicate a strong relationship, while values close to 0 indicate a weak relationship. Regression results in a regression equation (e.g., $Y=a+bX$), where a is the intercept, and b is the slope. The slope b tells us the rate at which the dependent variable changes for each unit change in the independent variable.

8.4 Types of Regression

Here is a look at the main types of regression techniques, each suited to specific types of data and analysis goals.

1. Simple Linear Regression

Simple Linear Regression is the most basic form of regression, examining the relationship between two variables—a dependent (response) variable and an independent (predictor) variable. The goal is to fit a straight line, described by the equation $Y=a+bX+\epsilon$. This type of regression is suitable when the relationship between the variables is approximately linear. For example, it might be used to predict sales based on advertising expenditure. The simplicity of this model makes it easy to interpret, but it's limited to linear relationships and only two variables.

2. Multiple Linear Regression

Multiple Linear Regression extends simple linear regression to include more than one independent variable, allowing for a more complex and realistic model. Its general equation is:

$$Y=a+b_1X_1+b_2X_2+b_3X_3+\cdots..+b_nX_n+\epsilon$$

Each X represents an independent variable, and each b is a coefficient showing the impact of that variable on Y. Multiple regression is useful in situations where the dependent variable is influenced by several factors. For instance, predicting house prices based on factors like location, square footage, number of bedrooms, and age of the property. However, as the number of variables increases, interpretation becomes more complex, and multicollinearity (high correlation between independent variables) can complicate the analysis.

3. Polynomial Regression

Polynomial Regression is a form of regression analysis where the relationship between the independent variable and the dependent variable is modeled as an n^{th}-degree polynomial.

Instead of a straight line, the regression curve is more flexible, taking shapes such as a parabola or a cubic curve. The general form for a polynomial regression model is: $Y=a+b_1X+b_2X^2+\cdots+b_nX^n$. This technique is suitable for data with non-linear relationships. For example, polynomial regression might model the growth of bacteria over time, where the relationship isn't linear but has a curved shape. While it can fit complex data patterns, polynomial regression may lead to overfitting if too many terms are included in the polynomial.

4. Logistic Regression

Despite its name, Logistic Regression is not used for linear relationships or predicting continuous outcomes. Instead, it is a classification technique used when the dependent variable is categorical—typically binary, with outcomes like "yes/no," "0/1," or "true/false." Logistic regression uses the logistic function, a sigmoid curve, to map predicted values to probabilities.

$$P(Y=1) = \frac{1}{1+e^{-(a+bx)}}$$

The predicted probability falls between 0 and 1, making this technique ideal for binary classification tasks. It is widely used in fields such as medicine for predicting disease presence or absence, and in finance for credit risk analysis. Logistic regression has also been extended to handle multiple classes (multinomial logistic regression) or ordered categories (ordinal logistic regression).

8.5 Simple Linear Regression

Simple Linear Regression is a foundational technique in statistical analysis, used to model the relationship between two continuous variables: an independent variable (predictor) and a dependent variable (response). It is one of the most widely used tools in data analysis, due to its simplicity and interpretability. The primary purpose of simple linear regression is to understand and quantify how the independent variable influences the dependent variable, and to make predictions.

In simple linear regression, the relationship between the independent variable X and the dependent variable Y is assumed to be linear. This means that the change in Y is proportional to the change in X, and it can be described by a straight line.

The general form of the simple linear regression model is:

$$Y=a+bX$$

Where:

- Y is the dependent variable (The main variable being studied or predicted).
- X is the independent variable (The variable used to predict or explain changes in the dependent variable).
- a is the y-intercept, representing the value of Y when $X= 0$
- b is the slope of the line, indicating the rate of change in Y for a one-unit increase in X.

Objectives of Simple Linear Regression

The main objectives of simple linear regression are:

1. To estimate the relationship between two variables: By analyzing the slope (b), we can determine the direction and strength of the relationship between X and Y.
2. To predict values: Using the regression line equation, we can predict the value of Y for any given value of X.
3. To quantify the goodness of fit: By evaluating the model's fit, we can assess how well it explains the variation in Y.

Formula of Regression Equation of Y on X

The general form of the regression equation of Y on X is expressed as:

$$Y=a+bX$$

Where:

- Y is the dependent variable we want to estimate based on X,
- X is the independent variable,
- a is the intercept of the regression line, and
- b is the slope or regression coefficient, representing the rate of change in Y for a unit change in X.

To find the values of a and b, the following two normal equations are derived from the least squares approach:

1. **Equation for Sum of Y Values:**

   $\sum Y = Na+b\sum X$

   where N is the total number of observations.

2. **Equation for Sum of XY Values:**

   $\sum XY = a\sum X+b\sum X^2$

Solving these simultaneous equations will provide the values of a and b, enabling the construction of the regression line of Y on X.

Illustration:

Calculate the regression equations of Y on X from the following data:

X	3	6	4	8	7
Y	4	7	5	10	9

Solution:

	X	Y	X^2	Y^2	XY
1	3	4	9	16	12
2	6	7	36	49	42
3	4	5	16	25	20
4	8	10	64	100	80
5	7	9	49	81	63
N=5	$\sum X=28$	$\sum Y=35$	$\sum X^2=174$	$\sum Y^2=271$	$\sum XY=217$

The equations are $\sum Y=Na+b\sum X$……..…..(i)

$$\sum XY=a\sum X+ b\sum X^2 \text{……..…..(ii)}$$

Substituting the values:

$35=5a+28b$ ……..(i)

$217=28a+174b$……..(ii)

Solving the equations (i) and (ii), we get,

Solve for a:

$5a=35-28b$

$a = \frac{35-28b}{5}$

Substitute this expression for *a* into Equation (ii),

$217=28(\frac{35-}{5}) +174b$

$217=\frac{980-784b}{5}+174b$

$1085=980-784b+870b$

$1085=980+86b$

$b = \frac{105}{86} = 1.22$

Now, substitute b=1.22 into the expression for *a*:

$a = \frac{35-28\times1.22}{5} = \frac{35-34.16}{5} = 0.168$

So, the regression line is $Y = 0.168+1.22X$.

Deviations taken from the Arithmetic mean of X and Y

When deviations are taken from the arithmetic mean of X and Y, the regression analysis process simplifies, especially in calculating the regression coefficients. This approach focuses on using the deviations from the mean values of the variables rather than the raw data itself, which can help in understanding the relationship between X and Y in a centered manner. Here's a breakdown of the steps and reasoning:

If X and Y are two variables with means $\overline{X}$ and $\overline{Y}$, we take deviations from these means as:

$$x = X-\overline{X} \text{ and } \quad y = Y-\overline{Y}$$

Here, *x* and *y* represent how much each value of X and Y deviates from their respective means.

With deviations from the mean, the regression equations of Y on X become:

$$Y-\overline{Y} = b_{yx}(X-\overline{X})$$

or, $y = b_{yx}\,.x$

Where, b_{yx} is the regression coefficient of Y on X.

With deviations from the mean, the regression equations of X on Y become:

$$X-\overline{X} = b_{yx}(Y-\overline{Y})$$

or, $x = b_{xy}\,.y$

Where, b_{xy} is the regression coefficient of X on Y.

The regression coefficients, when using deviations from the mean, are calculated as:

$$b_{yx} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} = \frac{\sum xy}{\sum x^2}$$

$$b_{xy} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(Y-\overline{Y})^2} = \frac{\sum xy}{\sum y^2}$$

Here:

$\sum xy$ is the sum of the product of deviations of X and Y.
$\sum x^2$ is the sum of the squares of deviations of X.
$\sum y^2$ is the sum of the squares of deviations of Y.

Illustration:

In the following table, the productivity scores of workers on a performance test and their monthly bonuses (in '000 Tk.) are recorded:

Worker	1	2	3	4	5	6	7	8	9	10
Productivity Score	35	55	45	65	70	50	60	75	40	55
Monthly Bonus ('000 Tk.)	3.0	5.5	4.0	6.0	6.5	4.5	5.5	7.0	3.5	5.0

Requirement:

a. Calculate the regression equation of the monthly bonus on the productivity score.
b. Estimate the probable monthly bonus for a worker with a productivity score of 85.

Solution: (a)

Worker	Productivity Score(X)	Monthly Bonus(Y)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
1	35	3.0	-20	-2.05	41.0	400
2	55	5.5	0	0.45	0.0	0
3	45	4.0	-10	-1.05	10.5	100
4	65	6.0	10	0.95	9.5	100
5	70	6.5	15	1.45	21.75	225
6	50	4.5	-5	-0.55	2.75	25
7	60	5.5	5	0.45	2.25	25
8	75	7.0	20	1.95	39	400
9	40	3.5	-15	-1.55	23.25	225
10	55	5.0	0	-0.05	0.0	0
N=10	$\sum X=550$	$\sum Y=50.5$			$\sum(X-\bar{X})(Y-\bar{Y})=150$	$\sum(X-\bar{X})^2=1500$

Mean of $\bar{X}=\frac{\sum X}{N}=\frac{550}{10}=55$, $\bar{Y}=\frac{\sum Y}{N}=\frac{50.5}{10}=5.05$

Regression Coefficient *b*

$b_{yx}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}=\frac{150}{1500}=0.10$

The Regression Line,

$Y-\bar{Y}=b_{yx}(X-\bar{X})$
$Y-5.05=0.10(X-55)$
$Y=0.10X-55\times0.10+5.05$
$Y=0.10X-5.5+5.05$
$Y=-0.45+0.1X$

(b) The Monthly Bonus for a Productivity Score of 85

Substitute X=85 into the regression equation:

$Y=-0.45+0.1\times85=-0.45+8.5=8.05$

The probable monthly bonus for a worker with a productivity score of 85 is approximately Tk.8.05 (in '000 Tk.)

Review Questions

1. What is the purpose of regression analysis?
2. Explain the difference between correlation and regression.
3. What is the regression equation, and how is it used?
4. How do you interpret the slope (b) in a regression equation?
5. Describe the difference between the regression of X on Y and the regression of Y on X.
6. What does it mean if the regression coefficient (b) is zero?
7. How is the line of best fit related to the regression equation?
8. Provide an example of a real-world situation where regression analysis could be useful.
9. Explain why deviations are often taken from the mean in regression calculations.
10. Discuss the importance of the regression line in predicting future values.
11. How does regression analysis help in identifying trends in data?
12. What is the formula for calculating the regression coefficient b?
13. How are deviations calculated from the mean in regression analysis?
14. Why do we use the least squares method in regression analysis?
15. Explain the assumptions made in simple linear regression analysis.
16. If a regression equation shows Y=3+0.5X, what does this imply about the relationship between X and Y?
17. A regression analysis results in the equation Y=10−0.2X. What does the negative slope tell us about the relationship between X and Y?
18. If the correlation between two variables is very high (e.g., 0.95), does it guarantee that one variable causes the other? Explain.
19. Compare and contrast the methods of calculating the regression line for Y on X and X on Y.
20. How does multiple regression differ from simple linear regression?
21. A study shows the following data for the years of experience (X) and monthly salary (Y) of employees:

Years of Experience (X)	4	7	8	10	11	13
Monthly Salary (Tk. in 000's) (Y)	20	25	27	30	33	36

Requirement:

a. Calculate the regression equation of Y on X.

b. Predict the monthly salary for an employee with 6 years of experience.

22. A teacher notices a pattern in the number of hours students spend watching TV per week (X) and their exam scores (Y):

Hours of TV	10	15	20	25	30	35	40	45	50	55
Exam Score	85	80	75	70	65	60	55	50	45	40

Calculate the regression equation of Y on X. Interpret the slope and what it suggests about the relationship between TV time and exam scores.

23. The Personnel Manager of a manufacturing company is looking for a way to determine the annual bonus (in thousands of Taka) for employees based on their years of experience. Data on the years of experience and corresponding annual bonuses from a sample of 10 employees is provided below:

Years of Experience (X)	3	6	8	5	9	10	4	7	2	11
Annual Bonus (Y) (Tk. '000)	8	12	15	11	16	18	10	14	7	20

Requirements:

a. Develop the regression equation of bonus Y based on the years of experience X.

b. Using the equation, estimate the annual bonus for an employee with 13 years of experience in a similar industry.

References

Berry, W., & Feldman, S. (1985). Multiple regression in practice. https://doi.org/10.2307/3151494.

Bin-Jie, J. (2004). On the Application of Regression Analysis in Economic Prediction.

Cameron, A., & Trivedi, P. (1998). Regression Analysis of Count Data. https://doi.org/10.2307/1271358.

Gunst, R., & Mason, R. (1980). Regression Analysis and Its Application: A Data-Oriented Approach. https://doi.org/10.2307/1267800.

Krishna, S., Oyebode, O., Joshi, A., Reddy, G., Karthikeyan, P., & Ganesh, K. (2023). The technical Role of Regression Analysis in prediction and decision making. *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 183-187. https://doi.org/10.1109/ICACITE57410.2023.10182872.

Lay, Y. (2009). Using multiple regressions in social sciences research: Some important aspects to be considered.