STA 111

LECTURE NOTE

Basic Statistics for Undergraduate

CHAPTER ONE

1.1 BRIEF HISTORY OF STATISTICS

The origin of statistics can be traced back to ancient civilizations, where it was primarily used for governmental and administrative purposes. Early evidence of statistical practice is found in records from ancient Egypt, China, and Mesopotamia, where population counts, taxation records, and agricultural data were collected.

1. **Ancient Roots**: The term "statistics" comes from the Latin word *"status"*, meaning "state" or "condition." In ancient times, rulers collected demographic and economic data to manage resources, plan for wars, and administer their territories. For example, the Roman Empire conducted censuses to manage tax systems.
2. **17th and 18th Centuries**: Modern statistics began to take shape in the 17th century with the rise of probability theory, thanks to mathematicians like Blaise Pascal and Pierre de Fermat. In the 18th century, statisticians like John Graunt and Edmond Halley applied statistical methods to demographic data, paving the way for actuarial science.
3. **19th Century Development**: The 19th century saw significant advances in statistical theory. Adolphe Quetelet introduced the concept of the "average man" in social statistics, and Carl Friedrich Gauss developed the normal distribution. Francis Galton introduced the concepts of correlation and regression.
4. **20th Century**: The foundations of modern inferential statistics were laid by mathematicians like Karl Pearson, who developed the chi-square test, and Ronald A. Fisher, who introduced concepts like analysis of variance (ANOVA) and maximum likelihood estimation. Fisher is often considered the father of modern statistics.

Today, statistics is an essential discipline across various fields such as economics, medicine, engineering, and the social sciences, with applications expanding rapidly in the era of big data and machine learning. decisions, it has several limitations:

1. **Data Dependence**:
   Statistics relies heavily on the quality and accuracy of the data. If the data is biased, incomplete, or inaccurate, the results of statistical analysis will be misleading. Poor data collection methods can introduce sampling errors and distort conclusions.
2. **Misuse of Statistical Methods**:
   Incorrect application of statistical techniques can lead to erroneous interpretations. For example, using the wrong type of test for a dataset or failing to account for confounding variables can produce misleading results.
3. **Assumptions in Statistical Models**:
   Many statistical methods are based on specific assumptions (e.g., normal distribution, independence of observations). If these assumptions are violated, the results may not be valid. Real-world data often does not fit perfectly into theoretical models, limiting the applicability of certain statistical techniques.
4. **Inability to Determine Causality**:
   Statistical analysis can identify correlations and associations between variables but cannot definitively establish causality. Even if two variables are statistically related, this does not mean one causes the other without further investigation, like experimental or longitudinal studies.
5. **Sensitivity to Sample Size**:
   Statistical significance is influenced by the size of the sample. Small sample sizes may lead to inaccurate results due to high variability, while large samples can make even trivial differences statistically significant. Care must be taken when interpreting results based on sample size.
6. **Over-Simplification**:
   Statistics often reduces complex phenomena into simplified numbers and models. This abstraction may overlook nuances and important details. A focus on averages and general trends might mask outliers or specific cases that are significant in practice.
7. **Potential for Misinterpretation**:
   Statistical results can be easily misinterpreted, either due to a lack of statistical literacy or intentional manipulation. For

example, a significant p-value might be misinterpreted as a large effect size, or correlation might be confused with causation.

8. **Ethical Concerns**:
   There is a potential for unethical use of statistics, such as manipulating data to support a particular agenda or misleading people with biased visual representations (e.g., deceptive graphs). Statistics should always be used transparently and responsibly.

DEFINITION

1.Statistics is the field of study that involves collecting, organizing, analyzing, interpreting, and presenting data. It helps to make decisions based on data, often dealing with uncertainty and variation.

2.

BRANCHES OF STATISTICS

- **Descriptive Statistics**: Summarizes or describes the characteristics of a data set. Examples include the mean (average), median (middle value), mode (most frequent value), standard deviation (a measure of variability).

  Example: In a classroom, the average score of students in a math test could be 70%, which is a descriptive statistic.

- **Inferential Statistics**: Makes inferences or predictions about a population based on data from a sample. This includes methods like hypothesis testing, confidence intervals, and regression analysis.

  Example: By surveying 200 patients in a hospital, you may infer the average blood pressure for the entire patient population.

Key Concepts:

- **Population**: The entire group of individuals or items that you want to study (e.g., all university students in Nigeria).
- **Sample**: A subset of the population used to represent the entire group (e.g., 200 university students sampled from various universities in Nigeria).
- **Parameter**: A numerical characteristic of a population (e.g., the average age of all university students in Nigeria).
- **Statistic**: A numerical characteristic of a sample (e.g., the average age of the 200 students sampled).

Case Study: Suppose you want to study the average income of families in a city. You could collect data from every family (which would be the population), but it may be expensive and time-consuming. Instead, you could randomly sample 100 families and use their data to estimate the average income for all families in the city.

1.2VARIOUS SOURCES OF STATISTICAL DATA

Statistical data can be collected from different sources, categorized as **primary** and **secondary** data.

1. **Primary Data**: Data collected firsthand by the researcher for a specific purpose.
   - **Methods of Collection**:
     - **Surveys**: Administering questionnaires to individuals either person to person or via the internet (mailed).
     - **Experiments**: Controlled tests to study variables.
     - **Observations**: Recording data based on direct observations.
     - **Interviews**: Conversations to gather information.

Example: Conducting a survey among 1,000 residents to understand the prevalence of diabetes in a community.

2. **Secondary Data**: Data collected by someone else for a different purpose but used by the researcher.

   - **Sources**:

     - **Government Reports**: Census data, national health surveys, economic reports.
     - **Academic Journals**: Peer-reviewed research papers.
     - **Administrative Records**: Hospital patient records, school attendance records.
     - **Online Databases**: United Nations, World Health Organization, World Bank, etc.

   Example: Using WHO data to analyze the spread of infectious diseases in West Africa.

Case Study: A research team is investigating the impact of air pollution on respiratory diseases. Instead of conducting new surveys, they use air quality data from government sources and patient data from hospitals (secondary data). They combine this data to determine whether there's a significant correlation between high levels of air pollution and increased cases of asthma.

1.3 IMPORTANT USES OF STATISTICS

Statistics is widely used across numerous fields for various purposes:

1. **Healthcare**:
   - Used in medical research and clinical trials to test the effectiveness of new drugs or treatments.
   - Public health agencies use statistics to track the spread of diseases and assess health interventions.

   Example: During the COVID-19 pandemic, statistical models were used to predict the spread of the virus and allocate medical resources accordingly.

2. **Business and Economics**:
   - Companies use statistics to make decisions about marketing strategies, product development, and customer behavior analysis.
   - Economists use statistical data to analyze trends like inflation, unemployment, and economic growth.

   Example: A retail store analyzes customer purchasing data to determine the most popular products during holiday seasons.

3. **Education**:
   - Teachers and educational institutions use statistics to evaluate student performance and develop teaching strategies.

   Example: A school might analyze students' exam scores to identify areas where teaching methods need improvement.

4. **Government Policy**:
   - Statistics help governments in planning infrastructure, health care, education, and social services by analyzing census data, employment rates, and other socioeconomic indicators.

   Example: Nigeria's government uses demographic statistics to plan for future educational and health services.

5. **Finance and Investment**
   - Statistical models are used to assess risks and returns on investment.

Example:

Case Study: A pharmaceutical company is testing a new drug for high blood pressure. They conduct clinical trials with 1,000 patients, using statistical methods to compare the effects of the new drug with a placebo. The analysis helps the company determine whether the drug is effective and safe for widespread use

1.4 USES OF STATISTICAL DATA

Statistical data is used in a variety of real-world applications:

1. **Forecasting**:
    - Used to predict future trends in sales, population growth, or weather patterns and so on

    Example: Businesses use sales data to forecast future demand for their products and so on.

2. **Quality Control**:
    - In industries, statistical data helps in monitoring production processes to ensure quality standards are met.

    Example: An automobile manufacturer uses data from product inspections to ensure cars meet safety standards.

3. **Risk Assessment**:
    - Banks and insurance companies use statistical data to assess risk and make decisions about lending or insuring clients.

    Example: An insurance company uses accident statistics to determine the premium rates for car insurance.

4. **Research and Development**:
    - Scientists and researchers use statistical data to test new hypotheses and develop new technologies.

1.5 QUANTITATIVE DATA

Example: A tech company uses data from customer feedback and product trials to improve its product design.

Case Study: A mobile phone company uses sales and demographic data to predict the future demand for smartphones in different regions. By analyzing this data, they adjust their production levels and marketing efforts to maximize sales.

Quantitative Data refers to data that can be measured and expressed numerically. It involves quantities and numbers and can be analyzed using mathematical techniques.

Types of Quantitative Data:

1. **Discrete Data**: Can only take certain values, often whole numbers.
    - **Example**: The number of students in a classroom (e.g., 25 students, and so on).
2. **Continuous Data**: Can take any value within a given range, and is measurable.
    - **Example**: A person's height, which could be 175.5 cm, or the temperature of a city on a particular day.

Examples of Quantitative Data:

- Number of malaria cases in a city.
- Weight of a sample of individuals.
- Income of a household in a given year, etc.

Case Study: A hospital collects data on the blood pressure of 500 patients over one month. This quantitative data is analyzed to

determine the average blood pressure and identify patterns among different age groups or genders. The data can be used to inform health interventions or patient care.

1.6 *VARIOUS SCALES OF MEASUREMENT*

Data can be classified based on different scales of measurement. Understanding these scales helps in selecting the appropriate statistical tools for analysis.

INTRODUCTION

In statistics, the concept of a "scale of measurement" refers to how we categorize and quantify data. Understanding the type of data, you're working with is crucial for choosing the appropriate statisticalmethods for analysis. There are four primary scales of measurement: **Nominal**, **Ordinal**, **Interval**, and **Ratio**. These scales differ in terms of the properties they possess (e.g., order, distance, or the presence of an absolute zero).

1. **Nominal Scale**

The nominal scale is the simplest form of measurement. Data on a nominal scale consist of categories or labels that do not imply any order or ranking. These categories are mutually exclusive and exhaustive, meaning that every observation falls into only one category, and all possible categories are listed.

- **Characteristics**:
  - No inherent order
  - Categories are qualitatively different, not quantitatively
  - Can only classify or count
- **Examples**:
  - Gender (Male, Female)
  - Blood Type (A, B, AB, O)
  - Marital Status (Single, Married, Divorced)
- **Permissible Operations**:
  - Counting frequencies
  - Mode (the most frequent category)
- **Statistical Tools**:
  - Chi-square tests
  - Proportions

2. **Ordinal Scale**

The ordinal scale represents a ranking or ordered relationship between categories. Although you can order the data, the differences between consecutive categories are not uniform or measurable. Ordinal scales only provide information on the relative position of an object in comparison to others.

- **Characteristics**:
  - Can be ranked or ordered
  - Intervals between ranks are not equal or defined
  - Non-numeric labels or rankings are common
- **Examples**:
  - Education Level (Primary, Secondary, Tertiary)
  - Pain Scale (No Pain, Mild Pain, Severe Pain)
  - Socioeconomic Status (Low, Middle, High)
- **Permissible Operations**:
  - Rank ordering
  - Median
  - Mode
- **Statistical Tools**:
  - Non-parametric tests (e.g., Mann-Whitney U, Kruskal-Wallis)
  - Percentile ranks

3. **Interval Scale**

The interval scale not only classifies and orders the data but also ensures that the intervals between values are meaningful and consistent. However, there is no absolute zero point, which means you cannot compute ratios (e.g., twice as much). This makes it possible to add and subtract values, but multiplication and division are not meaningful.

- **Characteristics**:
  - Equal intervals between points
  - No true zero point (zero does not mean "none")
  - Differences between values can be compared
- **Examples**:
  - Temperature (in Celsius or Fahrenheit)
  - IQ Scores
  - Calendar Dates
- **Permissible Operations**:
  - Addition and subtraction
  - Mean, median, and mode
- **Statistical Tools**:
  - Pearson correlation
  - t-tests, ANOVA

4. **Ratio Scale**

The ratio scale possesses all the characteristics of the interval scale, but with the added feature of an absolute zero point, which allows for the measurement of the absence of a variable. Because of the absolute zero, both differences and ratios between numbers are meaningful. This scale allows for the full range of mathematical operations: addition, subtraction, multiplication, and division.

- **Characteristics**:
  - True zero point (zero indicates the absence of the quantity)
  - Equal intervals between units
  - Ratios of measurements are meaningful
- **Examples**:
  - Height
  - Weight
  - Time
  - Age
- **Permissible Operations**:
  - All arithmetic operations: addition, subtraction, multiplication, and division
  - Mean, median, mode
- **Statistical Tools**:
  - Regression analysis
  - ANOVA
  - t-tests

K**ey Differences Between the Scales**

Feature	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Order	No	Yes	Yes	Yes
Equal Intervals	No	No	Yes	Yes
True Zero	No	No	No	Yes
Examples	Gender, Blood Type	Education Level, Pain Scale	Temperature, IQ Scores	Height, Weight, Age
Central Tendency	Mode	Median, Mode	Mean, Median, Mode	Mean, Median, Mode

Importance of Understanding Scales of Measurement

Understanding the scale of measurement for your data is crucial for several reasons:

- **Choice of Statistical Method**: Certain statistical techniques are only valid for specific types of data. For example, you can compute a mean for interval and ratio data, but not for nominal or ordinal data.
- **Validity of Results**: Applying inappropriate statistical methods can lead to misleading conclusions. For instance, running a Pearson correlation on ordinal data may not provide valid results.
- **Interpretation**: The interpretation of data differs based on the measurement scale. For instance, a difference in temperature measured on a Fahrenheit scale does not have the same implications as a difference in height measured in centimeters.

Case Study: A researcher is studying the effect of exercise on heart rate. Heart rate (measured in beats per minute) is ratio data since it has a meaningful zero point (no heartbeats) and the intervals between values are meaningful. This allows the researcher to calculate the percentage increase or decrease in heart rate after exercise.

CHAPTER TWO

2.1 BASIC SAMPLING TECHNIQUES

Sampling is the process of selecting a subset of individuals or observations from a population to estimate characteristics of the whole population. This helps when it's impractical or costly to collect data from every member of the population.

There are two main categories of sampling techniques: **probability sampling** and **non-probability sampling**.

A. PROBABILITY SAMPLING

In probability sampling, every member of the population has a known, non-zero chance of being selected. This approach is preferred when the goal is to make inferences about a population.

1.INTRODUCTION TO SIMPLE RANDOM SAMPLING

Simple Random Sampling (SRS) is a fundamental sampling technique in statistics where each member of a population has an equal chance of being selected, this can be done either by replacement or without replacement. It is the simplest and most straightforward method of probability sampling, allowing researchers to obtain a representative sample from a larger population. The key feature of SRS is its random nature, which minimizes bias and facilitates statistical inference.

Key Features of Simple Random Sampling

1. **Equal Probability of Selection**: Every individual in the population has the same chance of being included in the sample.

2. **Independence**: The selection of one individual does not influence the selection of another.
3. **Randomization**: The process involves random methods, which can be achieved through random number generators, lottery methods, or drawing lots.

Steps in Simple Random Sampling

1. **Define the Population**: Clearly specify the target population for the study. This can include a specific group of people, objects, or events.
2. **Determine Sample Size**: Decide how many individuals will be included in the sample based on research objectives and constraints.
3. **List the Population**: Create a complete list of all individuals in the population (sampling frame).
4. **Select the Sample**: Use a random method to select the sample from the population list.
5. **Collect Data**: Gather data from the selected individuals.
6. **Analyze Data**: Analyze the collected data to draw conclusions and make inferences about the population.

Example of Simple Random Sampling

Scenario: A researcher wants to study the eating habits of college students at a university with 1,000 students.

Step 1: Define the Population

- **Total Population**: 1,000 college students.

Step 2: Determine Sample Size

- **Sample Size**: The researcher decides to select 100 students for the study.

Step 3: List the Population

- Create a complete list of all 1,000 students, assigning each a unique identifier (e.g., Student ID).

Step 4: Select the Sample

- Use a random number generator or lottery method to select 100 unique student IDs from the list.

Step 5: Collect Data

- Administer a survey or questionnaire to the selected 100 students to gather data on their eating habits.

Total Sample Size

In this example, the total sample size is:

Sample Size=100 students

Summary of the Sampling Process

1. **Population**: 1,000 college students.
2. **Sample Size**: 100 students selected using simple random sampling.
3. **Data Collection**: Surveys administered to the selected students.

Advantages of Simple Random Sampling

- **Minimized Bias**: By giving every individual an equal chance of selection, SRS reduces the risk of bias in the sample.
- **Simplicity**: The method is straightforward to understand and implement, making it a popular choice for many researchers.

- **Statistical Validity**: The random nature of the sampling allows for valid statistical inferences to be made about the population.

Disadvantages of Simple Random Sampling

- **Requires a Complete List**: SRS necessitates a complete and accurate sampling frame, which may be difficult to obtain for large or dispersed populations.
- **Inefficient for Large Populations**: In very large populations, SRS can be time-consuming and resource-intensive.
- **Potential for Random Sampling Error**: Despite the random selection process, there is still the possibility of obtaining a non-representative sample purely by chance.

Applications of Simple Random Sampling

- **Surveys and Polls**: Widely used in market research and public opinion polls to ensure that the sample reflects the larger population.
- **Experimental Research**: Employed in experiments where researchers need a representative sample for testing hypotheses.
- **Epidemiological Studies**: Used in public health studies to assess health-related behaviors or outcomes across a population.

Exercises

1. **Exercise 1**: A researcher is conducting a study on the study habits of high school students in a district with 800 students. If they want to select a sample of 50 students, explain how they would implement simple random sampling.
2. **Exercise 2**: Discuss the advantages and disadvantages of simple random sampling compared to stratified random sampling.
3. **Exercise 3**: If the population size is 5,000 and the researcher wants a sample size of 200, calculate the sampling fraction.
4. **Exercise 4**: Explain how you would ensure that your sample is randomly selected using a random number generator

Simple Random Sampling is a fundamental method in statistics that provides a basis for obtaining a representative sample from a population. Its ease of use and ability to minimize bias make it a valuable tool for researchers. However, it is crucial to understand its limitations and to ensure that a complete sampling frame is available for effective implementation.

2. Introduction to Stratified Sampling

Stratified Sampling is a probability sampling technique used to obtain a representative sample from a population that is divided into distinct subgroups, known as strata. This method ensures that each subgroup is adequately represented in the sample, which enhances the precision of the estimates and the validity of the research findings. Stratified sampling is particularly useful when there are known differences between strata that may influence the variables being studied.

Key Features of Stratified Sampling

1. **Subgroup Identification**: The population is divided into homogeneous subgroups (strata) based on specific characteristics, such as age, gender, income, education level, etc.
2. **Proportional Representation**: Samples from each stratum are selected in proportion to the size of the stratum in the population.

3. **Increased Precision**: By ensuring that all relevant strata are represented, stratified sampling can reduce sampling error and provide more reliable results compared to simple random sampling.

Steps in Stratified Sampling

1. **Define the Population**: Clearly specify the target population for the study.
2. **Identify Strata**: Divide the population into relevant strata based on the characteristics of interest.
3. **Determine Sample Size**: Decide on the total sample size needed for the study.
4. **Select the Sampling Method for Each Stratum**: Choose how to sample within each stratum (e.g., simple random sampling, systematic sampling).
5. **Select Samples from Each Stratum**: Randomly select samples from each stratum according to their proportion in the population.
6. **Collect Data**: Gather data from the selected individuals.
7. **Analyze Data**: Analyze the data while considering the stratification design.

Example of Stratified Sampling

Scenario: A researcher wants to study the job satisfaction levels of employees in a large organization with 500 employees, divided into three departments: Sales, Marketing, and Operations.

Step 1: Define the Population

- **Total Population**: 500 employees.

Step 2: Identify Strata

- **Strata**:
  - Sales Department: 200 employees
  - Marketing Department: 150 employees
  - Operations Department: 150 employees

Step 3: Determine Sample Size

- **Total Sample Size**: The researcher decides to select 100 employees.

Step 4: Determine Proportions

- Calculate the proportion of employees in each stratum:
  - **Sales**:
  - **Marketing**: $\frac{n}{N} = \frac{150}{500} = 0.3$
  - **Operations**: $\frac{n}{N} = \frac{150}{500} = 0.3$

Step 5: Calculate Sample Sizes for Each Stratum

- **Sample Sizes**:
  - **Sales:** $\frac{n}{N} \times n = 0.4 \times 100 = 40$
  - **Marketing:** $\frac{n}{N} \times n = 0.3 \times 100 = 30$
  - **Operations:** $\frac{n}{N} \times n = 0.3 \times 100 = 30$

Step 6: Select Samples from Each Stratum

- Use simple random sampling to select 40 employees from Sales, 30 from Marketing, and 30 from Operations.

Step 7: Collect Data

- Administer a job satisfaction survey to the selected employees.

Summary of the Sampling Process

1. **Population**: 500 employees divided into 3 departments.
2. **Total Sample Size**: 100 employees.
3. **Strata and Sample Sizes**:
    - Sales: 40 employees
    - Marketing: 30 employees
    - Operations: 30 employees
4. **Data Collection**: Surveys administered to the selected employees.

Advantages of Stratified Sampling

- **Improved Accuracy**: By ensuring that all relevant subgroups are represented, stratified sampling increases the accuracy of estimates.
- **Reduction in Variability**: It reduces the overall variability of the sample estimates, leading to more precise results.
- **Specific Insights**: Enables researchers to analyze specific subgroups within the population, allowing for a more nuanced understanding of the data.

Disadvantages of Stratified Sampling

- **Complexity**: The process can be more complex than simple random sampling, requiring careful planning and execution.
- **Requires Detailed Population Information**: Accurate stratification necessitates detailed knowledge about the population and its characteristics.
- **Potential for Misclassification**: If individuals are incorrectly assigned to strata, it can introduce bias and affect the validity of the findings.

Applications of Stratified Sampling

- **Social Research**: Employed in studies that seek to understand different demographic groups, such as age, gender, or income levels.
- **Market Research**: Used to analyze consumer preferences across different market segments.
- **Health Studies**: Applied in public health research to ensure that various health subgroups are adequately represented.

EXERCISES

1. **Exercise 1**: A researcher is studying the academic performance of high school students in a district with 1,000 students divided into four grades (9th, 10th, 11th, 12th). If the researcher wants to select a sample of 200 students, describe how you would implement stratified sampling, including the calculations for each stratum.
2. **Exercise 2**: Discuss the advantages and disadvantages of stratified sampling compared to cluster sampling.
3. **Exercise 3**: If a researcher wants to study customer satisfaction across different store locations in a retail chain with 20 stores, how might they stratify their sample? Provide a rationale for your choices.
4. **Exercise 4**: Calculate the required sample sizes for each stratum if a researcher aims to sample a total of 500 individuals from a population of 2,000 divided into three strata with the following distributions: Stratum A (50%), Stratum B (30%), and Stratum C (20%).

Stratified Sampling is a powerful technique that enhances the representativeness and accuracy of survey results. By ensuring that all relevant subgroups are included in the sample, researchers can draw more valid conclusions about the population. Understanding

the intricacies of stratified sampling is essential for effective research design and data analysis.

3. Introduction to Systematic Sampling

Systematic Sampling is a probability sampling technique where researchers select samples from a population at regular intervals. This method involves selecting a starting point at random and then choosing every k^{th} individual from that point onward. Systematic sampling is often used for its simplicity and ease of implementation, particularly when a complete list of the population is available.

Key Features of Systematic Sampling

1. **Interval Selection**: Samples are selected at regular intervals (e.g., every 10th, 20th, etc.) after a random starting point.
2. **Structured Approach**: The systematic method provides a structured approach to sampling, making it straightforward to execute.
3. **Randomness**: Although the selection process follows a systematic pattern, the initial starting point is chosen randomly, maintaining the randomness of the sample.

Steps in Systematic Sampling

1. **Define the Population**: Clearly specify the target population for the study.
2. **Create a Sampling Frame**: Develop a complete list of all individuals in the population.
3. **Determine Sample Size**: Decide how many individuals you wish to include in the sample.
4. **Calculate the Sampling Interval**:
    - The sampling interval k is calculated as follows:

$$k = \frac{N}{n}$$

where $\boldsymbol{N}$ is the total population size, and $\boldsymbol{n}$ is the desired sample size.

5. **Select a Random Starting Point**: Randomly select a number between 1 and k as the starting point.
6. **Select the Sample**: From the starting point, select every k^{th}individual from the list until the sample size is reached.
7. **Collect Data**: Gather data from the selected individuals.
8. **Analyze Data**: Analyze the collected data to draw conclusions.

Example of Systematic Sampling

Scenario: A researcher wants to study the reading habits of library users in a town with a total of 1,000 registered library members.

Step 1: Define the Population

- **Total Population**: 1,000 library members.

Step 2: Create a Sampling Frame

- Create a complete list of all 1,000 members.

Step 3: Determine Sample Size

- **Sample Size**: The researcher decides to select 100 members.

Step 4: Calculate the Sampling Interval

$$k = \frac{N}{n} = \frac{1000}{100} = 10$$

Step 5: Select a Random Starting Point

- Assume the random number generator selects the number 7 as the starting point.

Step 6: Select the Sample

- Starting from member number 7, the researcher selects every 10th member:
  - 7, 17, 27, 37, 47, 57, 67, 77, 87, 97,
  - 107, 117, 127, 137, 147, 157, 167, 177, 187, 197,
  - 207, 217, 227, 237, 247, 257, 267, 277, 287, 297,
  - 307, 317, 327, 337, 347, 357, 367, 377, 387, 397,
  - 407, 417, 427, 437, 447, 457, 467, 477, 487, 497,
  - 507, 517, 527, 537, 547, 557, 567, 577, 587, 597,
  - 607, 617, 627, 637, 647, 657, 667, 677, 687, 697,
  - 707, 717, 727, 737, 747, 757, 767, 777, 787, 797,
  - 807, 817, 827, 837, 847, 857, 867, 877, 887, 897,
  - 907, 917, 927, 937, 947, 957, 967, 977, 987, 997.

Summary of the Sampling Process

1. **Population**: 1,000 library members.
2. **Sample Size**: 100 members selected using systematic sampling.
3. **Sampling Interval**: Every 10th member after starting from a randomly chosen member (number 7).
4. **Data Collection**: Surveys or assessments administered to the selected members.

Advantages of Systematic Sampling

- **Simplicity**: The process is straightforward and easy to implement, making it accessible for researchers.
- **Efficiency**: It can be quicker and more efficient than random sampling, particularly when dealing with large populations.
- **Evenly Distributed Sample**: Systematic sampling tends to yield a sample that is evenly distributed across the population.

Disadvantages of Systematic Sampling

- **Potential Bias**: If the population has a hidden pattern or periodicity, systematic sampling can introduce bias. For example, if every 10th individual has a particular characteristic, the sample may not be representative.
- **Not Suitable for Unknown Patterns**: If the population does not have a homogeneous structure, systematic sampling may not provide a valid representation.
- **Dependence on Sampling Frame**: It requires a complete and accurate list of the population, which may not always be available.

Applications of Systematic Sampling

- **Quality Control**: Often used in manufacturing and quality control processes to inspect products at regular intervals.
- **Surveys**: Applied in various surveys where a systematic approach is beneficial for data collection.
- **Epidemiological Studies**: Used in health studies to select participants based on systematic intervals for efficiency.

EXERCISES

1. **Exercise 1**: A researcher is studying consumer behavior in a mall with 2,000 shoppers. If they wish to sample 200 shoppers, describe how to implement systematic sampling and calculate the sampling interval.
2. **Exercise 2**: Discuss the differences between systematic sampling and simple random sampling, highlighting their respective advantages and disadvantages.
3. **Exercise 3**: If a researcher mistakenly selects a starting point that coincides with a periodic trend in the population, explain how this might impact the results.

4.**Exercise 4**: Calculate the sampling interval and select a sample of 50 from a population of 1,000 if the random starting point is determined to be 15.

4. Introduction to Multistage Sampling

Multistage sampling is a complex form of cluster sampling that involves selecting samples in multiple stages. Instead of selecting a sample from the entire population at once, multistage sampling divides the population into groups (or clusters) and then selects a sample of these clusters. Within each selected cluster, further sampling occurs to select the final subjects.

Key Features of Multistage Sampling

1.**Hierarchical Structure**: The sampling process occurs in stages, which can involve different sampling techniques at each stage (e.g., cluster sampling followed by simple random sampling).
2.**Cost-Effectiveness**: It is often more economical and practical than simple random sampling, especially in large populations spread over wide geographical areas.
3.**Flexibility**: Researchers can choose different sampling methods at different stages based on the needs of the study.

Steps in Multistage Sampling

1.**Define the Population**: Clearly specify the population you want to study.
2.**Select Primary Sampling Units (PSUs)**: Divide the population into clusters and randomly select a number of these clusters. These clusters are typically geographical areas.
3.**Select Secondary Sampling Units (SSUs)**: Within each selected cluster, perform further sampling to select individuals or smaller groups. This can involve stratified sampling, simple random sampling, or systematic sampling.
4.Data Collection: Collect data from the selected units.

Examples of Multistage Sampling

Example 1: Health Survey in Nigeria

1.**Population**: Adults in Nigeria.
2.**Primary Sampling Units**: Select a few states randomly (e.g., Lagos, Kano, and Rivers).
3.**Secondary Sampling Units**: Within each selected state, randomly choose several local government areas (LGAs).
4.**Tertiary Sampling Units**: Within each LGA, randomly select communities, and finally, choose individuals randomly within those communities for the survey.

Example 2:Educational Research

1.**Population**: High school students in a country.
2.**Primary Sampling Units**: Randomly select several provinces.
3.**Secondary Sampling Units**: Within each province, randomly choose schools.
4.**Tertiary Sampling Units**: Within each selected school, randomly select classrooms, and then randomly choose students from those classrooms.

Advantages of Multistage Sampling

- **Reduced Costs**: Less travel and administrative costs as clusters are often geographically concentrated.
- **Simplifies Data Collection**: Allows for manageable sampling sizes in large populations.
- **Flexibility in Methodology**: Different methods can be applied at each stage depending on the specific requirements of the study.

Disadvantages of Multistage Sampling

- **Complexity**: More complex to design and analyze than simple random sampling.

- **Sampling Error**: Greater risk of sampling error if the clusters are not representative of the population.
- **Bias Potential**: If clusters are not randomly chosen or are homogeneous, it can lead to bias in the results.

Exercises

1. **Exercise 1**: In a study of agricultural practices in Nigeria, outline the multistage sampling procedure you would use to select farmers from the population. Describe each stage in detail.
2. **Exercise 2**: Given a population of university students across various faculties in a country, design a multistage sampling strategy. Specify the sampling units at each stage and explain the rationale for your choices.
3. **Exercise 3**: Discuss how multistage sampling could be used to conduct a national health survey. Include considerations for sample size and how to ensure representativeness.
4. **Exercise 4**: Consider a scenario where you are researching the impact of education on health outcomes. Describe how you would implement multistage sampling to gather data from different educational institution.

Multistage sampling is a powerful technique for collecting data from large and diverse populations. By breaking the sampling process into stages, researchers can effectively manage resources while still achieving a representative sample. Understanding the principles and methodologies of multistage sampling is crucial for conducting effective research in various fields, including health, education, and social sciences.

Example

Scenario: A researcher wants to study the dietary habits of children in Nigeria. The population of interest includes children aged 5-12 years.

Step 1: Define the Population

- **Population**: Children aged 5-12 years in Nigeria.

Step 2: Select Primary Sampling Units (PSUs)

- **Clusters**: Nigeria is divided into 6 geopolitical zones.
- **Selection**: Randomly select 2 zones (e.g., North Central and Southwest).

Assumption: Each zone has a population of approximately 1,000,000 children aged 5-12 years.

Step 3: Select Secondary Sampling Units (SSUs)

- **Within Each Zone**: Divide each selected zone into Local Government Areas (LGAs).
- **Example**:
    - **North Central**: 10 LGAs
    - **Southwest**: 15 LGAs
- **Selection**: Randomly select 2 LGAs from each zone.

Selection:

- North Central: LGA1, LGA3
- Southwest: LGA5, LGA8

Step 4: Select Tertiary Sampling Units

- **Within Each Selected LGA**: Select communities.
- **Example**:
    - **LGA1**: 5 communities
    - **LGA3**: 4 communities
    - **LGA5**: 6 communities
    - **LGA8**: 5 communities
- **Selection**: Randomly select 1 community from each selected LGA.

Selection:

- From LGA1: Community A
- From LGA3: Community C
- From LGA5: Community E
- From LGA8: Community H

Step 5: Select Final Sample

- **Within Each Community**: Randomly select children.
- **Assumption**: Each community has approximately 200 children aged 5-12 years.
- **Selection**: Randomly select 10 children from each selected community.

Sample Selection:

- Community A: 10 children
- Community C: 10 children
- Community E: 10 children
- Community H: 10 children

Total Sample Size

The total number of children selected for the study is calculated as follows:

Total Sample Size=No. of Communities × Children per Community

Total Sample Size=4 Communities×10 Children=40 Children

Summary of the Sampling Process

1. **Primary Sampling Units**: 2 zones (North Central, Southwest)
2. **Secondary Sampling Units**: 4 LGAs (2 from each zone)
3. **Tertiary Sampling Units**: 4 communities (1 from each LGA)
4. **Final Sample**: 40 children selected (10 from each community)

Considerations for Analysis

1. **Data Collection**: The researcher will collect dietary data from the 40 selected children using a structured questionnaire.
2. **Analysis**: Data will be analyzed to assess dietary habits and their variations across the selected zones and communities.

Exercises Based on This Example

1. **Exercise 1**: If the researcher wanted to increase the sample size to 80 children, how many additional children would need to be selected from each community?
2. **Exercise 2**: If there were 5 communities in LGA1, how would you adjust your sampling strategy to ensure equal representation?
3. **Exercise 3**: Discuss potential sources of bias in this sampling strategy and how you would mitigate them.

5. Introduction to Cluster Sampling

Cluster Sampling is a probability sampling technique used when it is impractical or costly to conduct a simple random sample across a large population. Instead of sampling individuals directly, the population is divided into clusters (or groups), and entire clusters are randomly selected for inclusion in the sample. This method is particularly useful for geographical areas, large populations, and when a complete list of individuals is not available.

Key Features of Cluster Sampling

1. **Groups as Sampling Units**: The population is divided into clusters, which can be based on geographical areas, institutions, or any naturally occurring grouping.
2. **Random Selection of Clusters**: Instead of selecting individuals directly, entire clusters are randomly chosen.
3. **Cost-Effectiveness**: Reduces travel and administrative costs, making data collection more manageable and efficient.

4. **Two-Stage Sampling**: Often, cluster sampling can be a two-stage process where the first stage involves selecting clusters and the second stage involves sampling individuals within the selected clusters.

Steps in Cluster Sampling

1. **Define the Population**: Clearly specify the target population for the study.
2. **Identify Clusters**: Divide the population into clusters. Clusters should ideally be heterogeneous within and homogeneous between.
3. **Select Clusters**: Randomly select a number of clusters from the population.
4. **Select Individuals within Clusters**: Depending on the study design, either include all individuals from the selected clusters or randomly sample individuals within each selected cluster.
5. **Data Collection**: Collect data from the selected individuals.
6. **Analysis**: Analyze the data with consideration of the cluster sampling design.

Example of Cluster Sampling

Scenario: A researcher wants to study the academic performance of high school students across a country with a large number of schools.

Step 1: Define the Population

- **Total Population**: High school students in the country.

Step 2: Identify Clusters

- **Clusters**: The researcher divides the country into geographic regions or states. Assume there are 36 states in the country.

Step 3: Select Clusters

- **Random Selection**: The researcher randomly selects 5 states from the 36 states.

Step 4: Select Individuals within Clusters

- Within each selected state, the researcher randomly selects schools. Assume there are 10 high schools in each state.
    - Selected States:
        - State 1
        - State 2
        - State 3
        - State 4
        - State 5
- **Example Selection**:
    - **State 1**: Schools A, B, C
    - **State 2**: Schools D, E, F
    - **State 3**: Schools G, H, I
    - **State 4**: Schools J, K, L
    - **State 5**: Schools M, N, O
- The researcher decides to sample all students in the selected schools.

Step 5: Data Collection

- Surveys or assessments are administered to all students in the selected schools.

Total Sample Size

Assuming each selected school has an average of 200 students:

$$\textit{Total Sample Size} = \textit{Number of Schools} \times \textit{Average Number of Students per School}$$

$$\textit{Total Sample Size} = 15 \textit{ Schools} \times 200 \textit{ Students} = 3{,}000 \textit{ Students}$$

Summary of the Sampling Process

1. **Clusters Identified**: 36 states.
2. **Clusters Selected**: 5 states randomly chosen.
3. **Schools Selected**: 15 schools (3 schools from each state).
4. **Total Sample Size**: 3,000 students.

Advantages of Cluster Sampling

- **Cost-Effective**: Reduces travel costs and time, especially in geographically dispersed populations.
- **Simplicity**: Easier to administer and manage compared to simple random sampling in large populations.
- **Feasibility**: More practical when a complete list of ***individuals is unavailable.***

Disadvantages of Cluster Sampling

- **Increased Sampling Error**: Since the sampling occurs within clusters, there may be higher variability, leading to potential bias.
- **Limited Generalizability**: The findings may not be as generalizable to the entire population if clusters are not representative.
- **Homogeneity Within Clusters**: If clusters are too homogeneous, results may not reflect the diversity of the entire population.

Applications of Cluster Sampling

- **Public Health Studies**: Useful in epidemiological studies where geographic regions are sampled to assess health outcomes.
- **Education Research**: Employed to evaluate educational programs across different schools or districts.
- **Social Science Research**: Applied in surveys assessing social behaviors within communities.

Conclusion

Cluster sampling is a practical and efficient method for collecting data from large populations. By grouping the population into clusters and randomly selecting these clusters, researchers can save time and resources while still obtaining valuable data. However, understanding the potential limitations and biases associated with this technique is essential for drawing valid conclusions from the research findings.

Exercises

1. **Exercise 1**: A researcher wants to study the dietary habits of college students in a country with 200 universities. Describe how you would implement a cluster sampling strategy, including how you would select clusters and individuals within those clusters.
2. **Exercise 2**: Discuss the potential sources of bias in a cluster sampling design and how these might affect the study's conclusions.
3. **Exercise 3**: Compare cluster sampling with stratified random sampling. In which situations might one method be preferred over the other?
4. **Exercise 4**: If the researcher in the earlier example decided to randomly sample 10 students from each selected school instead of including all students, what would be the new sample size?

B. Non-Probability Sampling

In non-probability sampling, not all individuals have a chance of being selected. This approach is more convenient but limits the ability to generalize findings to the entire population.

1. **Convenience Sampling**:
   - The sample is selected based on availability and ease of access.
   - **Example**: A researcher stands in a mall and surveys anyone who walks by.

   Case Study: A researcher wants to study shopping habits and surveys customers entering a specific supermarket. This method is convenient but may not represent all shoppers.

2. **Purposive (Judgmental) Sampling**:
   - The researcher uses their judgment to select individuals who meet specific criteria.
   - **Example**: A health researcher selects individuals who have been diagnosed with a specific disease for a targeted study.

   Case Study: A nutritionist studying dietary habits of athletes could purposefully select professional athletes from various sports teams to gain insights.

3. **Snowball Sampling**:
   - Participants are asked to recommend other participants. This is commonly used in studies where finding participants is challenging.
   - **Example**: Researchers studying rare diseases may start with a few patients and then ask them to refer others who also have the disease.

   Case Study: A sociologist studying the experiences of immigrant workers may ask one participant to refer others, thus "snowballing" the sample.

4. Introduction to Quota Sampling

Quota Sampling is a non-probability sampling technique where the researcher ensures equal representation of specific subgroups within a population. This method is particularly useful when researchers want to gather data quickly or when the population is difficult to sample using probability techniques. Quota sampling focuses on achieving a specified number (quota) of participants from each identified subgroup, ensuring that these groups are represented in the final sample.

Key Features of Quota Sampling

1. **Non-Probability Sampling**: Participants are selected based on specific characteristics rather than random selection.
2. **Subgroup Representation**: The researcher defines quotas based on predetermined characteristics such as age, gender, ethnicity, income, etc.
3. **Flexibility**: Quota sampling allows researchers to adjust their approach based on the needs of the study or the availability of subjects.
4. **Cost-Effectiveness**: It can be more time-efficient and less expensive compared to probability sampling methods.

Steps in Quota Sampling

1. **Define the Population**: Clearly specify the target population for the study.
2. **Identify Quotas**: Determine the subgroups within the population that need to be represented (e.g., age groups, gender, ethnicity).
3. **Determine Sample Size**: Establish the total sample size needed for the study.
4. **Select Participants**: Recruit participants to meet the quotas for each subgroup. This can involve convenience sampling or judgment sampling methods.
5. **Data Collection**: Collect data from the selected participants.
6. **Analysis**: Analyze the data while taking into account the quotas set for each subgroup.

Numerical Example of Quota Sampling

Scenario: A researcher aims to study the reading habit of Kogi state Polytechnic students. The goal is to ensure representation across different grades and genders.

Step 1: Define the Population

- **Population**: Kogi state Polytechnic students with a total of 2,000 students.

Step 2: Determine Quotas

The researcher decides to establish quotas based on:

1. **Grade Level**: 9th, 10th, 11th, and 12th grades.
2. **Gender**: Male and Female.

Quota Breakdown:

- **Total Students**: 2,000
- **Grade Distribution**:
  - 9th Grade: 25% (500 students)
  - 10th Grade: 25% (500 students)
  - 11th Grade: 25% (500 students)
  - 12th Grade: 25% (500 students)
- **Gender Distribution**:
  - Males: 50% (1,000 students)
  - Females: 50% (1,000 students)

Total Quota Setup:

- **9th Grade**: 250 Males, 250 Females
- **10th Grade**: 250 Males, 250 Females
- **11th Grade**: 250 Males, 250 Females
- **12th Grade**: 250 Males, 250 Females

Total Quotas

- **9th Grade**: 500 (250 Males + 250 Females)
- **10th Grade**: 500 (250 Males + 250 Females)
- **11th Grade**: 500 (250 Males + 250 Females)
- **12th Grade**: 500 (250 Males + 250 Females)

Step 3: Select the Sample

The researcher approaches high schools in the city and collects data based on the established quotas.

- **Selection Process**:
  - The researcher goes to different schools and selects students until the quotas are filled.

Sample Selection:

1. **9th Grade**:

- Selected: 250 Males, 250 Females

2. **10th Grade**:
   - Selected: 250 Males, 250 Females
3. **11th Grade**:
   - Selected: 250 Males, 250 Females
4. **12th Grade**:
   - Selected: 250 Males, 250 Females

Total Sample Size

The total sample size is calculated as follows:

Total Sample Size=Quotas for Each Group

- **Total Quota per Grade**: 500 (from each grade level)
- **Total Sample Size**:

Total Sample Size=4 Grades×500 Students=2000 Students

Summary of the Sampling Process

1. **Quota Breakdown**:
   - 9th Grade: 250 Males, 250 Females
   - 10th Grade: 250 Males, 250 Females
   - 11th Grade: 250 Males, 250 Females
   - 12th Grade: 250 Males, 250 Females
2. **Total Sample Size**: 2000 students (500 from each grade with equal gender distribution).

Considerations for Analysis

1. **Data Collection**: The researcher collects data on reading habits through a questionnaire.
2. **Analysis**: Data will be analyzed to explore the differences in reading habits across grades and genders.

Advantages of Quota Sampling

- **Ensures Representation**: Guarantees that specific subgroups are adequately represented in the sample.
- **Quick and Cost-Effective**: Less time-consuming than probability sampling methods; useful for exploratory research.
- **Flexibility**: Researchers can modify quotas based on emerging insights during the study.

Disadvantages of Quota Sampling

- **Sampling Bias**: Since participants are not randomly selected, there is a risk of bias, leading to results that may not be generalizable to the entire population.
- **Lack of Randomization**: This method does not allow for statistical inference, limiting the ability to draw conclusions about the entire population based on the sample.
- **Subjective Selection**: The researcher's judgment in selecting participants can introduce bias.

Applications of Quota Sampling

- **Market Research**: Commonly used to assess consumer preferences, habits, and behaviors.
- **Public Opinion Polls**: Used to gather data on public sentiment regarding issues or events.
- **Social Research**: Employed in studies where specific demographic representation is crucial.

EXERCISES

1. **Exercise 1**: A researcher is studying the impact of social media on youth. If the population is divided into three age groups (13-17, 18-24, and 25-30), with a total sample size of

300, calculate the quotas for each age group assuming an equal representation.

2. **Exercise 2**: Discuss the potential biases that might arise in a study using quota sampling. How could these biases affect the research findings?
3. **Exercise 3**: Create a quota sampling plan for a study assessing the dietary habits of college students, including identified subgroups, total sample size, and quotas for each subgroup.
4. **Exercise 4**: Compare and contrast quota sampling with stratified random sampling. In which scenarios might one be preferred over the other

1. **Exercise 1**: If the researcher decides to increase the total sample size to 2,500 students, how should the quotas be adjusted?
2. **Exercise 3**: Discuss the potential biases associated with quota sampling and how they could affect the validity of the research findings.

Quota sampling is a valuable research method that allows for the collection of data from specific subgroups within a population. While it has its advantages in terms of efficiency and representation, researchers must be aware of the potential biases and limitations associated with this technique. Understanding when and how to apply quota sampling effectively is essential for conducting reliable research in various fields.

2.2 METHODS OF DATA COLLECTION

There are several methods of data collection, and choosing the right method depends on the research objectives, type of data needed, and available resources. Data collection methods can be divided into **primary** and **secondary** methods.

1. Primary Data Collection Methods

Primary data collection involves gathering new data specifically for the research problem at hand. It offers firsthand information but can be time-consuming and costly.

1.Surveys and Questionnaires:

- A structured set of questions is presented to participants to gather specific information.
- **Advantages**: Surveys can reach a large number of people, are cost-effective, and provide standardized data.
- **Disadvantages**: Response rates may be low, and responses could be biased due to poorly worded questions or respondent misunderstanding.

Example: A survey of 1,000 citizens to gather data on public opinion about healthcare services.

2.Interviews:

- Involves face-to-face, telephone, or online conversations where the researcher asks open-ended or structured questions.
- **Advantages**: Interviews provide in-depth information and clarification can be provided if respondents don't understand a question.
- **Disadvantages**: Time-consuming, expensive, and prone to interviewer bias.

Example: Conducting in-depth interviews with 50 farmers to understand the challenges they face with crop production.

3.Observations:

- The researcher collects data by directly observing the subjects or events.

- **Advantages**: Provides real-time and accurate data, especially for behavior studies.
- **Disadvantages**: Can be time-consuming, and subjects may alter their behavior when they know they are being observed.

Example: A researcher observes the shopping behavior of customers in a supermarket.

4.Experiments:

- The researcher manipulates one or more variables to observe the effect on a dependent variable.
- **Advantages**: Provides precise data on cause-and-effect relationships.
- **Disadvantages**: Often requires a controlled environment, which may not reflect real-world situations.

Example: A clinical trial where a new drug is tested on a group of patients to determine its effectiveness.

2. Secondary Data Collection Methods

Secondary data refers to data that was collected by someone else for a different purpose but is used by the researcher for their study. It saves time and resources but may not perfectly match the research objectives.

1.**Government Reports**:

- Reports and data sets published by government agencies like the census, labor statistics, health surveys, and economic data.
- **Advantages**: Reliable, comprehensive, and usually free to access.
- **Disadvantages**: May be outdated or lack specificity for the researcher's needs.

Example: Using the Nigerian census data to analyze population growth trends.

2.**Academic Journals and Research Papers**:

- Published studies that provide insights and data from previous research.
- **Advantages**: Peer-reviewed and generally reliable sources of data.
- **Disadvantages**: Access to some journals may be costly, and the data may not fully align with the current research needs.

Example: Reviewing published articles on the impact of climate change on agriculture.

3.**Administrative Records**:

- Data collected by institutions as part of their routine activities, such as hospital records, school enrollment data, or financial reports.
- **Advantages**: Often large-scale, comprehensive, and continuously updated.
- **Disadvantages**: Access may be restricted due to privacy laws, and data might be incomplete or inconsistent.

Example: Using hospital records to analyze the incidence of malaria in a specific region.

4.**Online Databases**:

- Data sets available on websites, including government portals, organizational reports, and statistical databases.

- **Advantages**: Easy access to a wide range of data.
- **Disadvantages**: The data might not be reliable or verified.

Example: Accessing World Bank data on global poverty trends.

Quota sampling is a non-probability sampling technique that allows researchers to ensure representation across specific subgroups within a population. While it can be useful for ensuring diversity, it also poses challenges related to bias and representativeness.

2.3 *DESIGN QUESTIONNAIRES AND FORMATS FOR DATA COLLECTION*

Introduction: A questionnaire is a research instrument consisting of a series of questions for the purpose of gathering information from respondents. The design of questionnaires is a critical step in the data collection process, as poorly designed questionnaires can lead to inaccurate data, misleading conclusions, and wasted resources.

Steps in Designing a Questionnaire:

1. **Define Objectives:**
   - Clearly define what you want to achieve from the data collection. This helps ensure that each question aligns with the objectives.
2. **Identify the Target Population:**
   - Consider who will be answering the questionnaire. The language and structure should be suitable for the target population.
3. **Choose the Type of Questionnaire:**
   - **Open-ended Questions:** Allow respondents to provide detailed answers.
   - **Close-ended Questions:** Provide fixed options for responses, such as multiple-choice, yes/no, or Likert scales.
4. **Draft the Questions:**
   - Ensure clarity and simplicity in wording.
   - Avoid leading or biased questions.
   - Maintain a logical flow, starting with general questions and moving to more specific ones.
   - Ensure questions are concise and relevant.
5. **Choose the Response Format:**
   - **Nominal scale:** Categorical responses without any order (e.g., gender, marital status).
   - **Ordinal scale:** Responses that indicate an order (e.g., satisfaction levels: very satisfied, satisfied, neutral, etc.).
   - **Interval/Ratio scale:** For numeric responses that have equal intervals (e.g., age, income, temperature).
6. **Pre-test the Questionnaire:**
   - Conduct a pilot test with a small group from the target population to identify any problems or areas of confusion.
7. **Finalize the Questionnaire:**
   - Based on the feedback from the pilot test, revise the questionnaire for clarity and accuracy.

Formats for Data Collection:

1. **Paper-Based:**
   - Traditional method of distributing questionnaires via printed forms.
   - Still used in areas with limited access to technology.
2. **Electronic (Online):**

- Can be distributed via email, social media, or survey platforms like Google Forms or SurveyMonkey.
- Enables quick data collection and analysis.

3. **Interviews:**
   - Structured or semi-structured interviews can be used, where the interviewer asks questions based on a questionnaire.
4. **Telephone Surveys:**
   - Useful for collecting data from respondents who may not have internet access or when face-to-face interactions are not possible.

Example 1: Survey Questionnaire

Purpose: To gather feedback on students' understanding of statistical concepts.

Title: Understanding of Key Statistical Concepts

Instructions: Please answer the following questions by ticking the most appropriate option or providing the required information.

1. **Gender:**
   - Male
   - Female
   - Other
2. **How comfortable are you with statistical data analysis?**
   - Very Comfortable
   - Comfortable
   - Neutral
   - Uncomfortable
   - Very Uncomfortable
3. **Which statistical software are you familiar with?** (Check all that apply)
   - SPSS
   - R
   - Python (Pandas, NumPy)
   - Excel
   - Other: _____
4. **On a scale of 1-5, how well do you understand the following statistical concepts?**
   - **Probability Theory**:
     - 1 (Poor)
     - 2 (Fair)
     - 3 (Good)
     - 4 (Very Good)
     - 5 (Excellent)
   - **Regression Analysis**:
     - 1 (Poor)
     - 2 (Fair)
     - 3 (Good)
     - 4 (Very Good)
     - 5 (Excellent)
   - **Hypothesis Testing**:
     - 1 (Poor)
     - 2 (Fair)
     - 3 (Good)
     - 4 (Very Good)
     - 5 (Excellent)

   - **Have you taken any formal course on statistics before?**

     - Yes
     - No

Example 2: **Survey Questionnaire**

Purpose: To gather data on programming language preferences among students.

Title: Programming Language Preferences Survey

Instructions: Please select the most appropriate option or provide information where necessary.

1. **Which programming languages are you currently proficient in?** (Check all that apply)
    - Python
    - Java
    - C++
    - JavaScript
    - Other: _____
2. **How many years have you been programming?**
    - Less than 1 year
    - 1-2 years
    - 3-5 years
    - More than 5 years
3. **Which area of Computer Science interests you the most?**
    - Web Development
    - Data Science
    - Machine Learning
    - Software Engineering
    - Cybersecurity
    - Other: _____
4. **On a scale of 1-5, how confident are you in the following programming skills?**
    - **Problem Solving**:
        - 1 (Not Confident)
        - 2
        - 3
        - 4
        - 5 (Very Confident)
    - **Debugging Code**:
        - 1 (Not Confident)
        - 2
        - 3
        - 4
        - 5 (Very Confident)
    - **Writing Algorithms**:
        - 1 (Not Confident)
        - 2
        - 3
        - 4
        - 5 (Very Confident)
5. **Do you prefer working on individual coding projects or team-based projects?**
    - Individual Projects
    - Team-Based Projects
    - No Preference

2.4 PROBLEMS AND TYPES OF ERRORS THAT ARISE IN DATA COLLECTION

Introduction: Data collection is the process of gathering and measuring information on variables of interest. However, this process can encounter several issues, leading to errors in the collected data. Identifying and mitigating these errors is crucial for ensuring the validity and reliability of the research results.

Types of Errors in Data Collection:

1. **Sampling Errors:**
    - These errors occur when the sample selected for the survey does not represent the entire population. Common causes include:
        - **Selection Bias:** When the method of selecting participants leads to a non-representative sample.
        - **Under-coverage:** When certain segments of the population are not included in the sample.
    - **Solution:** Use random sampling techniques to ensure that each member of the population has an equal chance of being selected.

2. **Non-Sampling Errors:**
    - **Measurement Error:** This occurs when there is a difference between the measured value and the true value. Causes include:
        - Poorly worded questions.
        - Misunderstanding by respondents.
        - Incorrect data entry.
    - **Solution:** Ensure clear, simple, and unambiguous questions. Train data collectors to minimize errors.
    - **Response Bias:** Occurs when respondents do not provide accurate answers due to social desirability, fear of judgment, or misinterpretation of the question.
    - **Solution:** Anonymity and confidentiality should be ensured. Use neutral language in questionnaires.
    - **Interviewer Bias:** Happens when the interviewer's presence, tone, or body language influences the respondent's answers.
    - **Solution:** Train interviewers to be neutral and objective.
    - **Non-Response Error:** Occurs when a significant portion of the sample does not respond to the survey. This can lead to biased results if the non-respondents differ significantly from respondents.
    - **Solution:** Follow up with non-respondents, offer incentives, or shorten the survey to encourage higher response rates.
3. **Data Processing Errors:**
    - These occur during the data entry, coding, and processing stages. Mistakes during transcription, incorrect coding, or errors during data analysis can lead to invalid conclusions.
    - **Solution:** Double-check data entry, use automated data processing systems where possible, and conduct thorough validation checks.
4. **Recall Bias:**
    - In surveys where respondents are asked to remember past events, they may provide inaccurate information due to memory decay or overestimation.
    - **Solution:** Use specific time frames in questions and minimize reliance on long-term memory.
5. **Instrument Error:**
    - Occurs when there are technical issues or inconsistencies in the tools used for data collection (e.g., a malfunctioning device or an unclear questionnaire).
    - **Solution:** Regularly calibrate instruments and test software before the data collection process.

Problems in Data Collection:

1. **Inaccessibility of Respondents:**
    - Some respondents may be difficult to reach due to geographical, technological, or social barriers.
    - **Solution:** Employ multiple methods for data collection (online, face-to-face, telephone) to improve accessibility.
2. **Time and Cost Constraints:**
    - Data collection can be time-consuming and expensive, particularly in large-scale studies.
    - **Solution:** Use cost-effective technologies (e.g., online surveys) and plan the study to maximize resource efficiency.
3. **Cultural and Language Barriers:**
    - In multicultural or multilingual populations, language differences may lead to misinterpretation of questions.
    - **Solution:** Translate questionnaires accurately and consider cultural sensitivities when designing questions.

Conclusion: Designing effective questionnaires and understanding the potential errors and challenges in data collection are key to ensuring the quality and reliability of research data. By employing rigorous design principles and mitigation strategies, researchers can minimize errors and improve the accuracy of their findings.

CHAPTER THREE

3.1 Distinguish Between Census and Sampling Surveys

Definition:

- **Census:** A complete enumeration of every individual in the entire population. Every unit is counted and surveyed.
- **Sampling Survey:** A process where only a subset of the population is selected and studied to infer conclusions about the entire population.

Key Differences:

- **Coverage:**
  - **Census** covers the entire population, while
  - **Sampling** covers only a portion or subset.
- **Cost and Time:**
  - **Census** is more expensive and time-consuming because it surveys everyone,
  - **Sampling** is cheaper and faster because it focuses on a smaller group.
- **Accuracy:**
  - **Census** gives more precise results because it accounts for everyone,
  - **Sampling** may have some sampling errors, but if well-done, it gives reliable estimates.
- **Feasibility:**
  - **Census** is feasible for small populations but can be impractical for large ones,
  - **Sampling** is more practical for large populations.

Example:

- **Census:** The Nigerian Population Census, where data is collected from every household.
- **Sampling Survey:** A survey of healthcare access among a randomly selected 1,000 Nigerians to infer the healthcare situation in the country.

EXERCISES:

1. Identify whether a census or sampling method is more suitable for studying:
   - Voter preferences in a national election
   - The effectiveness of a new drug for a rare disease

3.2 MEANING AND PURPOSE OF PILOT ENQUIRIES

Definition: A **Pilot Inquiry** (or pilot study) is a small-scale preliminary study conducted before the main survey or experiment. It is used to test the feasibility, time, cost, risk, and potential outcomes of a larger study.

Purpose:

- **Test Feasibility:** Ensures that the survey or study procedures work as planned.
- **Identify Issues:** Uncovers unforeseen problems, such as unclear questions in a survey or logistical problems.
- **Save Resources:** Reduces the risk of failure in the main study by testing the process on a smaller scale.
- **Improve Data Quality:** Helps refine the questionnaire, sampling techniques, and other methodologies.

Example: Before launching a nationwide survey on public health, a pilot study might be conducted in one small community to identify potential difficulties in data collection.

Exercise:

1. Why would you conduct a pilot inquiry before conducting a large-scale survey?
2. Imagine you are conducting a survey on internet usage habits. Design a small pilot study to test your survey.

3.3 Advantages and Disadvantages of Sampling

Advantages of Sampling:

- **Cost-Effective:** Sampling requires fewer resources (time, money, manpower) than conducting a census.
- **Faster Data Collection:** Results can be obtained quicker because fewer individuals are surveyed.
- **Less Burden:** It is less burdensome on respondents because fewer people need to be involved.
- **Flexibility:** Sampling allows for more in-depth questions or complex methodologies since fewer units are surveyed.

Disadvantages of Sampling:

- **Sampling Error:** There is a chance that the sample does not accurately represent the entire population, leading to biased results.
- **Non-Sampling Error:** Issues like poorly designed questions or incorrect data collection can lead to errors.
- **Limited Detail:** Sampling may not capture rare or extreme cases that are important but exist in small numbers.

Exercise:

1. List three scenarios where sampling would be better than a census.
2. How could you minimize sampling errors in a survey?

3.1INTRODUCTION TO POST ENUMERATION SURVEY (PES)

A Post Enumeration Survey (PES) is a survey conducted after a population census to evaluate the accuracy and completeness of the census data. PES helps identify undercounted or overcounted populations and assess the effectiveness of the census operation.

Objectives of PES:

- Assess the coverage of the census.
- Evaluate the accuracy of census data.
- Provide estimates of the population not counted during the census.
- Enhance the planning and execution of future censuses.

*2. **Importance of PES***

- **Quality Assurance:** PES is crucial for ensuring the reliability of census data, which is used for policy-making, resource allocation, and academic research.
- **Identifying Errors:** Helps identify systematic errors in census-taking procedures, such as non-response and misreporting.
- **Demographic Insights:** Provides insights into demographic trends and helps understand population dynamics.

*3. **Methodology of PES***

1. **Design Phase:**
   - Determine the survey objectives.
   - Select the sampling frame (usually based on the census enumeration areas).
   - Decide on the sample size and sampling method (probability sampling is often used).
2. **Implementation Phase:**
   - Conduct the survey using trained enumerators.
   - Collect data through various methods (interviews, online surveys, etc.).
   - Ensure that the PES is conducted in a manner that is independent of the census operation.
3. **Analysis Phase:**
   - Compare PES results with census data.
   - Calculate coverage rates, undercounting rates, and overcounting rates.
   - Analyze demographic characteristics of the surveyed population.

Examples of PES

Example 1: United States 2010 Census PES

- The U.S. Census Bureau conducted a PES to assess the accuracy of the 2010 Census. The survey revealed that approximately 1.5% of the total U.S. population was undercounted. The results were used to adjust the census figures for demographic analyses.

Example 2: Nigeria's 2006 Population Census PES

- Following the 2006 census, Nigeria conducted a PES that highlighted significant undercounting in rural areas. This information was crucial for subsequent policy adjustments and resource allocations.

EXERCISES

Exercise 1: Coverage Calculation

- In a hypothetical PES, a sample of 1,000 households was surveyed, and it was found that 950 of them were also counted in the census. Calculate the coverage rate.

Solution: Coverage Rate = (Number of Households counted in both the PES and the Census / Total Households surveyed) × 100 = (950 / 1000) × 100 = 95%

Exercise 2: Under-coverage Estimation

- In a PES conducted in a region, it was estimated that 30,000 individuals were missed in the census. If the total population according to the census is 1,000,000, what is the under-coverage rate?

Solution: Under-coverage Rate = $\frac{Estimated\ Missed\ Population}{Total\ Population\ according\ to\ Census} \times 100$

$$= \left(\frac{30{,}000}{1{,}000{,}000}\right) \times 100 = 3\%$$

Exercise 3: Data Analysis Interpretation

- Review the following hypothetical results from a PES:
    - Total individuals surveyed: 2,000
    - Individuals reported in census: 1,800
    - Individuals not reported in census: 200

Interpret the results and discuss potential reasons for the discrepancies.

Suggested Interpretation:

- The PES indicates that 200 individuals were not reported in the census, leading to an undercounting rate of 10%. Potential reasons for this discrepancy could include:
    - Non-response in specific demographic groups.
    - Errors in data collection during the census.
    - Migration patterns affecting the population count.

Conclusion

Post Enumeration Surveys are essential for improving census operations and ensuring accurate population data. The insights gained from PES inform future census planning, demographic research, and policy-making.

CHAPTER FOUR

4.1 DATA COLLECTION, CLASSIFICATION, VERIFICATION, STORAGE, AND COMPILATION

1. Introduction to Data

Data serves as the foundation for analysis, research, and decision-making across various fields. Understanding how to collect, classify, verify, store, and compile data is essential for ensuring accurate and meaningful insights.

Definition of Key Terms:

- **Data:** Raw facts and figures that can be processed to generate information.
- **Information:** Processed data that is meaningful and useful for decision-making.
- **Statistics:** A branch of mathematics dealing with data collection, analysis, interpretation, presentation, and organization.

2. Categories of Collected Data

Data can be classified into different categories based on its nature, scale, and measurement. Understanding these categories helps in determining the appropriate methods for analysis.

A. Quantitative Data

Quantitative data is numerical and can be measured or counted. It is further divided into two main types:

1. **Discrete Data:**
    - **Definition:** Countable data that can take specific values (e.g., whole numbers).
    - **Examples:**
        - Number of students in a class (e.g., 25).
        - Number of cars in a parking lot (e.g., 15).
        - Scores on a test (e.g., 85, 90, 92).
2. **Continuous Data:**
    - **Definition:** Measurable data that can take any value within a range.
    - **Examples:**
        - Height of individuals (e.g., 160.5 cm, 175.2 cm).
        - Temperature readings (e.g., 20.5°C, 30.2°C).
        - Time taken to complete a task (e.g., 1.5 hours).

B. Qualitative Data

Qualitative data is non-numerical and describes characteristics or qualities. It is divided into two main types:

1. **Nominal Data:**
    - **Definition:** Categorical data that cannot be ranked or ordered.
    - **Examples:**
        - Colors of cars (red, blue, green).
        - Types of fruit (apple, banana, cherry).
        - Gender (male, female, non-binary).
2. **Ordinal Data:**
    - **Definition:** Categorical data that can be ranked or ordered, but the intervals between ranks are not uniform.
    - **Examples:**
        - Customer satisfaction ratings (poor, fair, good, excellent).
        - Education levels (high school, bachelor's, master's).
        - Survey responses (disagree, neutral, agree).

3. Classification of Data

Data classification involves organizing data into defined categories for easier analysis and interpretation. This step is crucial for understanding data structure and preparing for further analysis.

A. Classifying Quantitative Data

1. **Discrete Data Example:**

- **Scenario:** A survey collects data on the number of pets owned by individuals in a neighborhood.
- **Data Set:** {0, 1, 1, 2, 3, 3, 4}
- **Classification:**
  - Count occurrences of each value:
    - 0 pets: 1
    - 1 pet: 2
    - 2 pets: 1
    - 3 pets: 2
    - 4 pets: 1

2.**Continuous Data Example:**

- **Scenario:** A study measures the height of individuals in a population.
- **Data Set:** {160.2, 162.4, 165.1, 168.5, 170.0}
- **Classification:**
  - Data can be grouped into ranges for analysis, e.g.,
    - 160-165 cm
    - 166-170 cm

B. Classifying Qualitative Data

1.**Nominal Data Example:**

- **Scenario:** Data on types of transportation used by respondents.
- **Data Set:** {car, bus, bike, car, bus, train}
- **Classification:**
  - Count occurrences of each type:
    - Car: 2
    - Bus: 2
    - Bike: 1
    - Train: 1

2.**Ordinal Data Example:**

- **Scenario:** Customer satisfaction survey with ratings.
- **Data Set:** {good, excellent, fair, poor, excellent}
- **Classification:**
  - Rank the responses:
    - Poor: 1
    - Fair: 2
    - Good: 3
    - Excellent: 4

4. Verification of Sorted Data

Verification is a critical step in the data management process, ensuring that the data collected and classified is accurate and reliable. This process helps identify errors or discrepancies that may affect analysis and decision-making.

A. Methods of Verification

1.**Cross-Validation:**

- Compare the sorted data with the original data set.
- Example: If a survey indicates 120 respondents but the original list shows only 115, further investigation is needed.

2.**Statistical Techniques:**

- **Consistency Checks:** Verify that similar data points are consistent across different data collections.
- **Range Checks:** Ensure that data points fall within expected ranges (e.g., ages should not be negative).
- **Outlier Detection:** Identify and investigate data points that fall outside the expected range.

3.**Data Audits:**

- Conduct systematic checks of data entries and classifications.
- Example: Randomly sample data points to ensure accuracy and consistency.

B. Example of Verification Process

- **Scenario:** After sorting a data set of students' grades.
- **Sorted List:** {75, 80, 80, 85, 85, 90, 95}
- **Original Data Set:** {80, 85, 90, 75, 85, 80, 95}
- **Verification Steps:**
    - Compare both lists to ensure all grades are included.
    - Count duplicates to ensure they are reflected accurately in the sorted list.

Exercise 1: Verification

Given the sorted list of ages: {22, 25, 25, 30, 30, 30, 35}, verify against the following original data set: {25, 30, 30, 22, 35, 30, 25, 30}. Identify any discrepancies.

5. DATA STORAGE METHODS

Data storage is essential for maintaining the integrity and accessibility of data. Different storage methods can be employed based on the type and volume of data.

A. File-Based Storage

- **Description:** Data is stored in files on a disk or storage device (e.g., CSV, TXT, Excel).
- **Pros:**
    - Simple and easy to access.
    - Ideal for small to medium-sized data sets.
- **Cons:**
    - Not ideal for complex queries or large data sets.
    - Potential issues with data integrity and security.

B. Database Management Systems (DBMS)

- **Description:** Use of software applications to create and manage databases (e.g., MySQL, PostgreSQL, Oracle).
- **Pros:**
    - Efficient for large data sets.
    - Supports complex queries and data relationships (using SQL).
- **Cons:**
    - Requires technical expertise to manage and maintain.
    - Potentially higher setup and maintenance costs.

C. Cloud Storage

- **Description:** Data stored on remote servers accessed via the internet (e.g., Google Drive, AWS S3, Microsoft Azure).
- **Pros:**
    - Scalable and flexible; can accommodate varying data sizes.
    - Accessible from anywhere with an internet connection.
    - Often includes built-in backup solutions.
- **Cons:**
    - Privacy and security concerns regarding data access and ownership.
    - Dependence on internet connectivity.

D. Data Warehousing

- **Description:** Centralized storage for large amounts of historical data designed for analysis (e.g., Amazon Redshift, Snowflake).
- **Pros:**
    - Optimized for reporting and analysis.

- Integrates data from multiple sources for a comprehensive view.

- **Cons:**
  - Expensive to set up and maintain.
  - Complexity in data migration and integration.

6. COMPILATION OF DISCRETE AND CONTINUOUS DATA

Data compilation is the process of organizing and summarizing data for analysis. This step is crucial for deriving meaningful insights from the collected data.

A. Compiling Discrete Data

- **Example:** A survey collects data on the number of pets owned by individuals.
- **Data Set:** {0, 1, 2, 3, 3, 4, 5}

Compilation Steps:

1. **Frequency Distribution:**
   - Count occurrences of each value:
     - 0 pets: 1
     - 1 pet: 1
     - 2 pets: 1
     - 3 pets: 2
     - 4 pets: 1
     - 5 pets: 1
2. **Visualization:**
   - Create a bar chart to represent the frequency of pets owned.
   - Each bar represents the count of individuals owning a certain number of pets.

B. Compiling Continuous Data

- **Example:** A study measures the height of individuals in a population.
- **Data Set:** {160.2, 162.4, 165.1, 168.5, 170.0}

Compilation Steps:

1. **Class Intervals:**
   - Create ranges for analysis:
     - 160-165 cm
     - 166-170 cm
2. **Frequency Distribution:**
   - Count occurrences within each interval:
     - 160-165 cm: 3
     - 166-170 cm: 2
3. **Visualization:**
   - Create a histogram to represent the distribution of heights.
   - Each bar represents the number of individuals within a specific height range.

Exercise 2: Compilation

1. Given the following discrete data set of exam scores: {78, 85, 90, 78, 92, 85, 85, 90}, compile a frequency distribution table and draw a bar chart.
2. For the following continuous data set of weights (in kg): {55.2, 60.0, 62.5, 65.0, 70.5}, create class intervals, compile a frequency distribution, and draw a histogram.

Understanding the various categories of data, methods of classification, verification processes, storage techniques, and compilation of discrete and continuous data is crucial for effective

data analysis. Proper data management ensures accurate results, informed decision-making, and meaningful insights.

EXERCISES

Exercise 1:
Create a frequency table for the following data representing the scores of 20 students in a test:
72, 65, 80, 82, 70, 95, 60, 75, 85, 90, 68, 72, 70, 62, 80, 78, 85, 92, 68, 60.

Exercise 2:
Using the contingency table below, answer the following:

- Calculate the row and column totals.
- What is the relationship between the two variables (gender and course enrollment)?

	Science	Arts	Commerce
Male	15	10	5
Female	12	18	10

Exercise 3:
Plot a scatter diagram using the following data on the relationship between hours studied and test scores:

Hours Studied	Test Score
1	50
2	55
3	65
4	70
5	80

Exercise 4:
Discuss the advantages and disadvantages of using a pie chart versus a bar chart to represent data.

CHAPTER FIVE

5.1THE VARIOUS TYPES OF STATISTICAL TABLES

Statistical tables provide an organized way to present data. Different types of tables serve different purposes in statistics.

1. Frequency Tables: A frequency table lists the number of times each value or group of values appears in a dataset. They are useful in summarizing categorical or grouped data.

- **Example:**
  Test scores of 30 students:

Test Score Range	Frequency
0-10	2
11-20	5
21-30	10

Test Score Range	Frequency
31-40	8
41-50	5

- **Use:** Frequency tables are great for quickly identifying the distribution of values in a dataset.

2. Contingency Tables: A contingency table, also known as a cross-tabulation or crosstab, shows the distribution of two or more categorical variables. They are often used to study the relationship between variables.

- **Example:** Relationship between gender and smoking habits among university students:

	Smoker	Non-Smoker	Total
Male	30	70	**100**
Female	10	90	**100**
Total	**40**	**160**	**200**

- **Use:** Contingency tables are commonly used in chi-square tests to test independence between variables.

3. Simple Informative Tables: These tables present summarized information in a clear, easy-to-understand format. They are often used in descriptive statistics.

- **Example:** Monthly sales of a

Month	Sales (Units)
January	120
February	150
March	200

retail store:

4. Reference Tables: These tables are used to refer to statistical values like z-scores, t-distributions, and chi-square values. These values help in hypothesis testing and confidence interval calculations.

- **Example:** Z-score table for standard normal distribution.

5. Complex Tables: Complex tables display large amounts of information, often with multiple rows and columns, providing detailed comparisons.

- **Example:** A table showing the performance of students in three subjects across five different semesters:

Student Name	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5
John	75	80	78	82	85
Mary	68	72	70	75	78
Ahmed	82	85	80	84	88

- **Use:** Complex tables are typically used in longitudinal studies, economic data, and cross-sectional analysis.

5.2 METHODS OF DATA PRESENTATION

Data can be presented in several forms, including tabular, graphical, and pictorial

1. **Tabular Presentation:** This method involves organizing data into rows and columns in a table format. It is an effective way to summarize numerical information

formats. Each method has its advantages depending on the nature of the data.

- **Example:** Monthly rainfall in a year:

Month	Rainfall (mm)
January	50
February	40
March	80
April	100

- **Advantages:**
  - Easy to summarize large amounts of data.
  - Can show detailed information.

Disadvantages:

  - Not very visual.
  - May be overwhelming for large datasets.

2. Graphical Presentation: Graphs are visual representations of data that help in understanding trends, distributions, and comparisons more easily.

- **Types of Graphs:**
  - **Bar Chart:** A bar chart displays categorical data with rectangular bars. The length of each bar is proportional to the value it represents.

  - **Histogram:**

Example: Sales of different products:

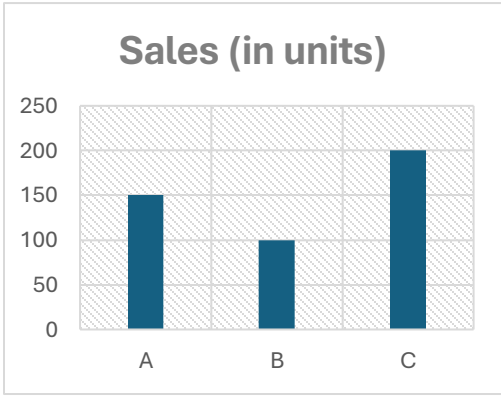


Fig 1 A Bar Chat

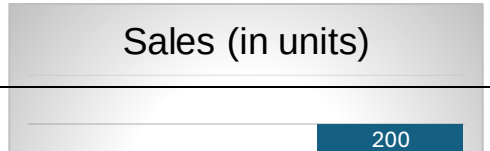


Fig 2: A Histogram

BAR CHART:

Advantages:

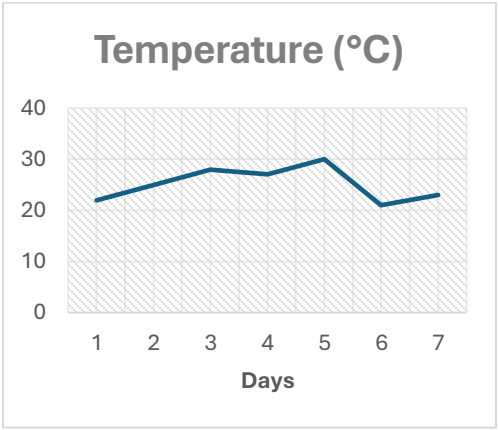
- Useful for comparing categories.
- Easy to understand.

Disadvantages:

- Not effective for showing trends over time.

- **Line Graph:** Line graphs are used to track changes over periods of time. A line graph connects data points with straight lines.

  Example: The changes in temperature over the past 7 days:



LINE GRAPH:

Advantages:

- Effective for showing trends over time.
- Simple to interpret.

Disadvantages:

- Not suitable for categorical data.

- **Pie Chart:** A pie chart divides data into slices, where each slice represents a proportion of the total.

  Example: Market share of different companies:

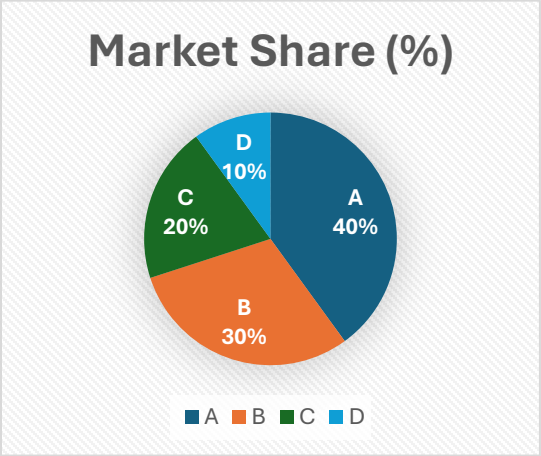


Fig 3: A Pie Chart

PIE CHART:

Advantages:

- Good for representing parts of a whole.
- Visually appealing.

Disadvantages:

- Difficult to compare multiple pie charts.
- Not suitable for large datasets.

3. Pictorial Presentation: Pictograms use symbols or images to represent data values. Each symbol corresponds to a specific value.

- **Example:** Number of cars sold by a dealership using car icons to represent units.

Advantages:

- Highly visual and easy to interpret.

Disadvantages:

- Less accurate than other methods, especially for precise data.

4. Textual Presentation: In textual presentation, data is described using words. This method is often used when describing data trends, insights, or summaries.

5.3 CONSTRUCT SCATTER DIAGRAMS, FREQUENCY TABLES, AND GRAPHS

1. **Scatter Diagram:** A scatter plot is used to investigate the relationship between two numerical variables. Each point on the graph represents one observation.

- **Example:**
  Investigating the relationship between hours studied and test scores:

Hours Studied	Test Score
2	50
3	60
4	70
5	80

- **Scatter Plot:**
- **Use:** Scatter diagrams help identify correlations (positive, negative, or none) between variables.

2. Frequency Tables:
These tables display how often different values occur within a dataset (see Section 5.1 for examples).

3. Graphs (See Section 5.2):
Use graphs such as histograms, line graphs, bar charts, and pie charts to visually represent data.

5.4 EXPLAIN THE MERITS AND DEMERITS OF CHARTS/DIAGRAMS

BAR CHART:

- **Merits:**
  - Easy to construct and interpret.
  - Can handle both positive and negative values.
  - Effective for comparing quantities across categories.
- **Demerits:**
  - May mislead if scales are not uniform.
  - Not suitable for continuous data.

PIE CHART:

- **Merits:**
  - Best for representing parts of a whole.
  - Visually appealing and easy to understand.
- **Demerits:**
  - Difficult to compare data between different pie charts.
  - Less effective for datasets with many categories.

LINE GRAPH:

- **Merits:**
  - Best for showing trends over time.
  - Easy to interpret.
- **Demerits:**
  - Not suitable for representing categorical data.
  - Requires careful scaling.

SCATTER DIAGRAM:

- **Merits:**
  - Shows relationships between variables.
  - Highlights outliers and trends.

- **Demerits:**
    - Can be difficult to interpret without knowledge of regression analysis.
    - Does not work well with large datasets.

5.5 LIFE DATA

Life data refers to the time to an event of interest, such as the failure of a machine or the death of a patient. This is used extensively in survival analysis and reliability studies.

Example:
A medical study tracking the survival of cancer patients after a particular treatment. The data includes the survival time (in months) of patients and whether they experienced the event (death) during the study period.

Time (Months)	Number at Risk	Number of Events
1	100	5
2	95	10
3	85	7

Survival Curve: A Kaplan-Meier curve is used to estimate the survival probability over time.

EXERCISES

Exercise 1:
Create a frequency table for the following data representing the weights (in kg) of 30 people:
50, 55, 60, 60, 65, 70, 75, 50, 55, 60, 70, 85, 90, 60, 65, 70, 75, 85, 60, 70, 90, 55, 60, 65, 70, 75, 85, 50, 60, 70.

Exercise 2:
Using the contingency table below, answer the following questions:

- Calculate the total row and column frequencies.
- Analyze the relationship between age and car ownership.

Age Group	Own Car	Don't Own Car	Total
18-25	10	30	
26-35	25	15	
36-45	20	10	

Exercise 3:
Plot a scatter diagram using the following data on the relationship between advertising spending (in dollars) and sales (in thousands):

Advertising ($)	Sales (000s)
500	25
700	30

Advertising ($)	Sales (000s)
1000	40
1200	45
1500	50

Exercise 4:
Discuss the merits and demerits of using a line graph versus a scatter plot to represent the relationship between time spent on social media and test scores.

MEASUREMENT OF CENTRAL TENDENCY

A. UNGROUP DATA

1. MEAN

Example:

Consider the following dataset of scores:

12, 15, 20, 15, 22, 25, 30, 15, 10, 18

Mean: The mean is the sum of all values divided by the number of values.

$$\text{Mean } (\bar{x}) = \frac{12+15+20+15+22+25+30+15+10+18}{10} = \frac{182}{10} = 18.2$$

2. MEDIAN

To find median, first, arrange the data in ascending order:

10, 12, 15, 15, 15, 18, 20, 22, 25, 30

Since there are 10 numbers (even number), the median is the average of the 5th and 6th numbers:

$$\text{Median} = \frac{15+18}{2} = 16.5$$

3. MODE

The mode is the value that appears most frequently in the dataset. In this case the number 15 appears three times, so:

Mode= 15

B. GROUPED

Consider the following frequency distribution of test scores:

SCORES (CLASS INTERVAL)	FREQUENCY
10 – 19	5
20 – 29	8
30 – 39	12
40 – 49	7
50 – 59	3

MEAN:

Step 1: Calculate the midpoints ($\frac{uci+lci}{2}$) of each class interval.

Class Interval	Midpoint (x)	Frequency (f)	fx
10 – 19	**14.5**	**5**	**72.5**
20 – 29	**24.5**	**8**	**196.0**

30 – 39	**34.5**	**12**	**414.0**
40 – 49	**44.5**	**7**	**311.5**
50 – 59	**54.5**	**3**	**163.5**

Step 2: Calculate the sum of the products $\boldsymbol{fx}$and the total frequency $\sum \boldsymbol{f}$:

$$\sum \boldsymbol{fx} = 72.5 + 196.0 + 414.0 + 311.5 + 163.5 = 1157.5$$

$$\sum \boldsymbol{f} = 5 + 8 + 12 + 7 + 3 = 35$$

Step 3: Compute the mean$(\bar{x})$:

$$\textbf{Mean}(\bar{\boldsymbol{x}}) = \frac{\sum \boldsymbol{fx}}{\sum \boldsymbol{f}} = \frac{1157.5}{35} = 33.07$$

MEDIAN

Step 1: Identify the median class

The total frequency is 35. So, $\frac{\mathbf{35}}{\mathbf{2}} = \mathbf{17.5}$. The class where the cumulative frequency reaches or exceeds 17.5 is the median class, which is $\mathbf{30 - 39.}$

Step 2**:** Use the formula for median:

$$\textbf{Mean} = \boldsymbol{L} + \left(\frac{\frac{n}{2} - F}{f_m}\right) \times \boldsymbol{h}$$

Where:

$\boldsymbol{L} = 29.5$ (Lower boundary of the median class)

$\boldsymbol{F} = \mathbf{13}$(Cumulative frequency before the median class)

$\boldsymbol{f_m} = \mathbf{12}$(Frequency of the median class)

$h = \mathbf{10}$(Class width)

Median = $\mathbf{29.5} + \left(\frac{\mathbf{17.5-13}}{\mathbf{12}}\right) \times \mathbf{10} = \mathbf{29.5} + \left(\frac{\mathbf{4.5}}{\mathbf{12}}\right) \times \mathbf{10} = \mathbf{29.5 + 3.75 = 33.23}$

MODE

Use the mode formula for grouped data:

Mode =$\boldsymbol{L} + (\frac{\boldsymbol{f_m - f_1}}{\boldsymbol{2f_m - f_1 - f_2}}) \times \boldsymbol{h}$

Where:

$\boldsymbol{L} = \mathbf{29.5}$ (Lower class boundary of the modal class, which is also **30-39**)

$\boldsymbol{f_m} = \mathbf{12}$ (Frequency of the modal class)

$\boldsymbol{f_1} = \mathbf{8}$ (Frequency of the class before the modal class)

$\boldsymbol{f_2} = \mathbf{7}$ (The frequency of the class after the modal class)

$h = \mathbf{10}$(Class width)

Mode = $\mathbf{29.5} + \left(\frac{\mathbf{12-8}}{\mathbf{2(12)-8-7}}\right) \times \mathbf{10} = \mathbf{29.5} + \left(\frac{\mathbf{4}}{\mathbf{9}}\right) \times \mathbf{10} = \mathbf{29.5 + 4.44 = 33.94}$

MEASUREMENT OF DISPERSION

A. UNGROUP

RANGE, VARIANCE AND STANDARD DEVIATION

Example:

Consider the dataset:

12, 15, 20, 15, 22, 25, 30, 15, 10, 18

RANDGE

The range is the difference between the maximum and minimum values.

Range = $30 - 10 = 20$

VARIANCE AND STANDARD DEVIATION

Step 1: Compute the mean (as found above: $\bar{x} = \mathbf{18.2}$)

Step 2: Calculate the squared deviations from each value from the mean:

$(12 - 18.2)^2 = 38.44; (15 - 18.2)^2 = 10.24;$

$(20 - 18.2)^2 = 3.24; (15 - 18.2)^2 = 10.24$

$(22 - 18.2)^2 = 14.44; (25 - 18.2)^2 = 46.24;$

$(30 - 18.2)^2 = 139.24; (15 - 18.2)^2 = 10.24;$

$(10 - 18.2)^2 = 67.24; (18 - 18.2)^2 = 0.04$

Step 3: Calculate the variance (for a sample, divide by $n - 1$, and for a population, divide by n)

Assuming this is a sample:

Variance = $\frac{38.44+10.24+3.24+10.24+14.44+46.24+139.24+10.24+67.24+0.04}{10-1} = \frac{339.6}{9} = 37.73$

Step 4: Compute the standard deviation as the square root of the variance

Standard Deviation (S.D) = $\sqrt{37.73} \approx 6.14$

INTER QUARTILE RANGE (IQR)

The IQR is the difference between the third quartile (Q_3) and the first quartile (Q_1).

Class interval	Midpoint(x)	Frequency(f)	fx	fx^2
10 – 19	14.5	5	72.5	1051.25
20 – 29	24.5	8	196.0	4802.0
30 – 39	34.5	12	414.0	14338.5
40 – 49	44.5	7	311.5	13837.75
50 – 59	54.5	3	163.5	8918.25

Step 1: Arrange the data in ascending order:

10, 12, 15, 15, 15, 18, 20, 22, 25, 30

Step 2: (Q_1) is the median of the lower half:

(Q_1) = Median of $[10, 12, 15, 15, 15] = 15$

Step 3: (Q_3) is the median of the upper half

(Q_3) = median of $[18, 20, 22, 25, 30] = 22$

Step 4: The interquartile range:

$$\boldsymbol{IQR = Q_3 - Q_1 = 22 - 15 = 7}$$

NB: In a situation where the total number of data points is odd, this means the middle (median) value after arrangement in ascending order is Q_2 while the values before Q_2 is the lower half and the values after Q_2 is the upper half.

B. GROUPED DATA

VARIANCE, STANDARD DEVIATION AND SEMIINTERQUATILE RANGE

i. VARIANCE AND STANDARD

Example: Using the frequency table:

Class interval	Frequency (f)
10 – 19	4
20 – 29	6
30 – 39	9
40 – 49	5
50 - 59	3

Class interval	Frequency (f)	Cumulative frequency (cf)
10 – 19	4	4
20 – 29	6	10
30 – 39	9	19
40 – 49	5	24
50 – 59	3	27

Step 1: Compute $\sum fx$ and $\sum fx^2$:

$$\sum fx = 1157.5\ ; \qquad \sum fx^2 = 42947.75$$

Step 2: Compute the variance using the formula:

$$\text{Variance} = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

$$\text{Variance} = \frac{42947.75}{35} - \left(\frac{1157.5}{35}\right)^2 = 1227.08 - (33.07)^2 = 1227.08 - 1093.62 = 133.46$$

Step 3: Compute the standard deviation:

$$\text{Standard Deviation (SD)} = \sqrt{133.46} \approx 11.55$$

MEASUREMENT OF DISPERSION- COEFFICIENT OF VARIATION (CV)

The coefficient of variation is a normalized measure of dispersion, calculated as:

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$\text{CV} = \frac{11.55}{33.07} \times 100 \approx 34.92\%$$

ii. INTERQUARTILE RANGE

Example: Consider the following grouped data representing test scores:

To

find the interquartile range, we need to calculate the cumulative frequencies and identify the first quartile (Q_1) and third quartile (Q_3) values.

Identify N and Quartile Positions:

$N = 27$ (Total frequency)

Position of $Q_1 = \frac{N}{4} = \frac{27}{4} = 6.75$

Position of $Q_3 = \frac{3N}{4} = \frac{3(27)}{4} = 20.25$

Locate Q_1:

The 6.75^{th} value falls withing the cumulative frequency that reaches or exceeds 6.75, which is the second-class interval (20-29).

Using the Q_1 formula:

$$\boldsymbol{Q_1 = L + (\frac{\frac{N}{4} - F}{f}) \times h}$$

Where:

L = 19.5 (Lower boundary of the 20-29 class)

F = 4 (Cumulative frequency before this class)

f = 6 (Frequency of this class)

h = 10 (class with)

$Q_1 = 19.5 + \left(\frac{6.75-4}{6}\right) \times 10 = 19.5 + (\frac{2.75}{6}) \times 10$

$Q_1 = 19.5 + 4.58 = 24.08$

Locate Q_3:

The 20.25^{th} value falls within the cumulative frequency that reaches or exceeds 20.25, which is the fourth-class interval (40-49).

Using the Q_1 formula:

$$\boldsymbol{Q_3 = L + (\frac{\frac{3N}{4} - F}{f}) \times h}$$

Where:

L = 39.5 (Lower boundary of the 40 - 49 class)

F = 19 (Cumulative frequency before this class)

f = 5 (Frequency of this class)

h = 10 (class with)

$Q_3 = 39.5 + \left(\frac{20.25-19}{5}\right) \times 10 = 39.5 + (\frac{1.25}{5}) \times 10$

$Q_3 = 39.5 + 2.5 = 42.0$

Calculate the Interquartile Range (IQR)

IQR = $Q_3 - Q_1 = 42.0 - 24.08 \approx 17.92$

CENTRAL TENDENCIES

1. Mean

- **Advantages**:
    - Uses all data points, providing a comprehensive measure.
    - Suitable for further statistical analysis (e.g., t-tests).
- **Disadvantages**:

- Sensitive to extreme values (outliers).
- Not suitable for skewed distributions.

2. Median

- **Advantages**:
  - Less affected by outliers and skewed data.
  - Good measure of central tendency for ordinal data.
- **Disadvantages**:
  - Ignores most of the data except for the middle value.
  - Not ideal for further statistical analysis.

3. Mode

- **Advantages**:
  - Useful for categorical data.
  - Easy to identify in a distribution.
- **Disadvantages**:
  - Not useful for continuous data.
  - May not exist or may have multiple modes, complicating interpretation.

MEASUREMENT OF DISPERSION

1. Range

- **Advantages**:
  - Simple to calculate.
  - Gives a quick sense of the spread.
- **Disadvantages**:
  - Affected by extreme values.
  - Ignores all data points except the two extremes.

2. Variance and Standard Deviation

- **Advantages**:
  - Uses all data points, providing a comprehensive measure.
  - Useful for further statistical calculations and analyses.
- **Disadvantages**:
  - Sensitive to outliers.
  - May be challenging to interpret in the context of the original units.

3. Interquartile Range (IQR)

- **Advantages**:
  - Not affected by extreme values, as it focuses on the middle 50% of data.
  - Useful for skewed data.
- **Disadvantages**:
  - Ignores the spread of data outside the middle 50%.
  - Less informative for symmetric distributions.

CHAPTER SIX

PROBABILITY, SET THEORY AND PROBABILITY DISTRIBUTION

The likelihood of chance that en event may likely occur (thought not certain) is known as probability of that event. In other words, it is the quantitative measure of uncertainty of an event. To therefore measure any prevailing uncertainty accurately. Probability techniques are employed.

6.1 Occurrence of chance of Events

Here are the occurrence of chance of events:

(i). The chance of a head of Events when a fair coin is tossed.

(ii). The chance of 1 showing up when a fair die is rolled.

(iii). The chance of a particular team winning a football match in a penalty shoot-out.

(iv). The chance of hitting a target when a stone is thrown at a flying bird.

(v). The chance of picking a card from a well shuffled pack of playing cards.

The outcome of any of these events is left to chance and we therefore base our

decision on degree of rational belief.

6.2 Types of probability

Types of probability are:

(i). Theoretical Probability

This probability is based on information known about a physical situation. It is

an expected probability void of experiment and it occurs when every event has

equal chance of occurring. Theoretical probability of an event is defined as

$$P(E) = \frac{\text{Number of favourable outcome}}{\text{Total number of possible outcome}}$$

NOTE: Possible outcome consist of favourable and unfavourable outcome.

(b). Empirical or Experimental Probability

This probability involves carrying out an experiment. It is

not an exact or final

probability as further trials can definitely affect the result.

Different people carrying out the same experiment may

obtain different results. If a die is

tossed a number of times and we are interested in an event, then the empirical

probability of such an event is defined as:

$$P(E) = \frac{\text{Number of times the event occurs}}{\text{Total number of trials}}$$

6.3 Properties of Probability Measure

The properties of probability measure are as follow:

(i). The probability of an event E ranges from 0 to 1 inclusive, i.e $0 \leq P(E) \leq 1$.

This implies that we neither have negative probability nor probability greater than 1 (i.e probability of any event is between 0 and 1 inclusive).

(ii). The probability of uncertainty (impossibility) is 0. A value of 0 indicates that there is no chance that the favourable event will occur. For Illustration, the probability that one will be able to run non-stop from Abuja to Lagos is 0.

(iii). The sum of probability of all sample points in a sample space equal to 1.

(iv). If the probability that an event will occur is P, then probability that it will not occur is $p' = 1 - p$

6.4 Basic Concepts of Probability

The basic concept of probability are:

(a). **Any experiment:** Any process that generate data and its outcome cannot be predicted with certainty.

(b). **Probability Experiment:** The experiment has more than one possible outcome and the outcome depends on chance. For Illustration, when a fair coin is tossed,

a head or tail appears but the one that will appear cannot be determined in advance.

(c). **Sample space:** This is the set of all the possible outcomes in a random experiment. When a coin is tossed once, the sample space is $\{H, T\}$ and when a die is tossed once, the sample space is {1,2,3,4,5,6}.

(d). **Sample point:** This is an element of a sample space. In the sample space for a tossed die, 1 is a sample point and each of the remaining outcomes.

(e). **An event:** This is a subset of a sample space (one or more sample points). Each of the sets $\{13\}, \{2,4\}, \{2,4,6\}$ is an event.

(f). **Equally likely events:** Two or more events are said to be equally likely possible if any of them cannot be expected to occur in preference to the others (i.e they have equal chance of occurrence).

Illustration 8.1

In a class of 30 students, 18 are boys and the remaining students are girls.

What is the probability of selecting a girl from the class at random?

Solution

Total number of students = 30

Number of boys = 18

Number of girls = 30-18 = 12

Probability of selecting a girl $= \frac{12}{30} = \frac{2}{5}$

Illustration8.2

A fair dice is tossed twice, what is the probability of obtaining a total of 5?

Solution

A die has 6 faces and the sample space is generated below

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

Total number of possible outcome = 36

Number of required event n(E) = {1,4},{2,3}, {3,2}, {4,1}

Probability (obtaining a total of 5) = $\frac{n(E)}{n(s)} = \frac{4}{36} = \frac{1}{9}$

Illustration 8.3

If a fair coin is tossed three times. What is the probability of getting:

(a). 3 heads (ii). no head (iii). at least 1 head

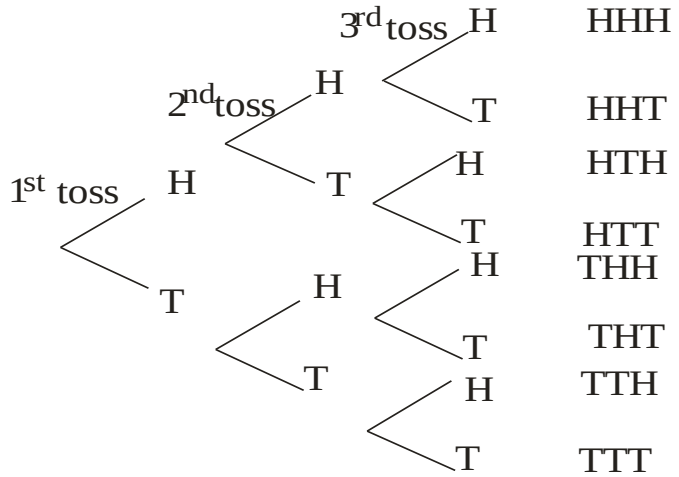
(iv). at least 2 heads (v). 2 heads

Solution

When a coin is tossed, then we have $S = \{H, T\}$ that is two outcome

Using probability tree diagram, we have 8 outcomes when the coin is tossed

three times:



S = $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

n(s) = 8

(i). Prob. (getting 3 heads) $= \frac{1}{8}$ i.e E = $\{HHH\}, n(E) = 1$

(ii). Prob (getting no head) = $\frac{1}{8}$ i.e E = $\{TTT\}, n(E) = 1$

(iii). Prob (getting at least 1 head) =

$\frac{7}{8}\{HHH.HTH, HTT, THH, TTH, THT, HHT\}$, $n(E) = 7$

(iv). Prob (getting at least 2 heads) = $\frac{4}{8} = \frac{1}{2}$

(v). Prob (getting 2 heads) = $\frac{3}{8}$ i.e E = $\{HTH, THH, HHT\}$

n(E) = 3

6.5 Types of Events

Events has different types which are as follow:

(i). **Mutually Exclusive Events**

Two events A and B are mutually exclusive if the occurrence of A precludes the occurrence of B. In other words, the two events cannot occur together. For Illustration, when a coin is tossed once, a head or a tail cannot appear at the same time. For two mutually exclusive events E_1 and E_2 $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ that is there is no intersection between events E_1 and E_2.

For n mutually exclusive events, $E_1, E_2, \ldots E_n$.

$P(E_1 \cup E_2 \cup \ldots .E_n) = P(E_1) + P(E_2) + \ldots + P(E_n)$

(ii). Non-Mutually Exclusive Events

Two events are non-mutually exclusive when they can occur at the same time.

For Illustration, tossing of a coin and rolling of a die are two mutually exclusiveevents.

(iii) **Dependent Events**

Two events are dependent if the occurrence of one is connected in some way with the occurrence of the other. For Illustration, in the results of drawing two cards from a pack, one at a time without replacement, the result of the second card depends on the first. When the first card is drawn, there are only 51 cards left from which we have to draw the second card.

(iv) Independent Events

Two events are independent if the occurrence of one event is not connected in any way with the occurrence of the other. The occurrence of event A does not depend on whether B has occurred or not. For Illustration, when a die is cast twice, the result of the second toss is independent of the result of the first toss.

8.6 Basic Laws of Probability

The basic laws of probability are:

(i). Addition rule for mutual exclusive events

The probability of the union of events E_1 and E_2 where E_1 and E_2 are mutually exclusive events is given as $P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$.

(ii). Addition rule for non-mutually exclusive events

The probability of the unions of the events E_1 and E_2 where E_1 and E_2 are non-

mutually exclusive events is given as $P(E_1 \text{ or } E_2) = P(E_1)+P(E_2) - (E_1 \cap E_2)$.

(iii). Multiplication rule for independent events

Given events E_1 and E_2 which are independent, the probability of their

simultaneous occurrence is given as:

$P(E_1 \text{ and } E_2) = (E_1 \cap E_2) = P(E_1).P(E_2)$

(iv). Multiplication rule for dependent events

Given events E_1 and E_2 (which are jointly dependent) their joint probability is

given as:

$P(E_1 \text{ and } E_2) = P(E_1). P(E_2).P(E_2/E_1)$

$P(E_2/E_1)$ depicts the conditional probability of E_2 given that E_1 has occurred

$$P(E_2/E_1) = \frac{P(E_2 \cap E_1)}{P(E_1)}$$

Illustration8.4

Find the probability that a single toss of a die will result in:

(i). a number greater than 3

(ii). an odd number

(iii). a prime number

Assume equal probability for the sample point

Solution

Table 6.2

X	1	2	3	4	5	6
P(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

(i). Let E_1 be the event of numbers greater than 3

$E_1 = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

(ii). Let E_2 be the event of odd numbers

$E_2 = \{1,3,5\}$

$P(E_2) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

(iii). Let E_3 be the event of prime numbers

$E_3 = \{2,3,5\}$

$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

Illustration 8.5

Four balls were drawn at random from a box containing 4 red, 5 blue, 7 white

and 3 yellow balls. Find the probability that they were drawn in the order: red,

blue, white and yellow if each ball is:

(i). replaced

(ii). not replaced

Solution

(a). Let R,B,W,Y stand for Red, Blue, White and Yellow balls respectively. n(R) = 4, n(B) = 5, n(W) = 7, n(Y) = 3, n(S) = 19.

Since there is sampling with replacement, the events are independent. The

required probability is:

$$P(R \cap B) \cap (W \cap Y) = P(R).P(B).P(W).P(Y)$$

$$= \frac{4}{19} \times \frac{5}{19} \times \frac{7}{19} \times \frac{3}{19} = \frac{420}{130321} = 0.0032228$$

(b). This is sampling without replacement. Thus, the events are dependent. The required probability is:

$$P(R \cap B \cap w \cap Y) = P(R).P(B/R).P(W/R \cap B)$$

$$P(Y/R \cap B \cap W) = \frac{4}{19} \times \frac{5}{18} \times \frac{7}{17} \times \frac{3}{16} = \frac{420}{93024} =$$

0.0045153

CHAPTER SEVEN

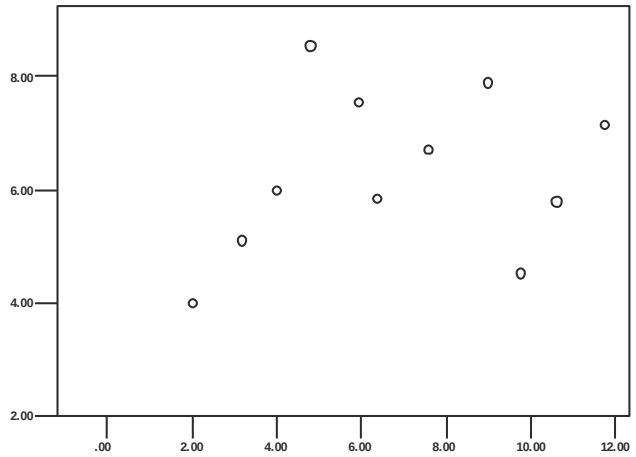
REGRESSION ANALYSIS

In many problems, there are two or more variables that are related and it is necessary to explain the nature of this relationship. For example, in a study that involves ability to save of some civil servants in Kogi state, this may be related to some variable (factors) such as level of education, income, family size etc. it may be of interest to build a model related to the ability to save of the civil servants and use the model for prediction.

Suppose there is a single dependent variables which depends on k independent variables, e.g. x1,x2K, xx. The relationship between these variables is characterized by a mathematical model called a regression equation.

SCATTER DIAGRAM

Scatter diagram is a graph representing two series with known variable to be plotted on x-axis and the variable to be estimated plotted on the y-axis. Example of a scatter diagram is shown bellow.



METHOD OF OBTAINING A REGRESSION LINE

There are 3 methods used to obtaining a regression line

(i) Use of scatter diagram or freehand method

(ii) Semi average method

(iii) The least spares method

Free hand method

This is the simplest method of determining the regression line. The idea is to draw a line on a scatter diagram by a free hand in such a way that it passes through the centre of the points. It is done by taking the mean value of x and that of y and ensures that the regression line passes through this point (x, y).

The main disadvantage of this is that different people would probably draw different line using the same data.

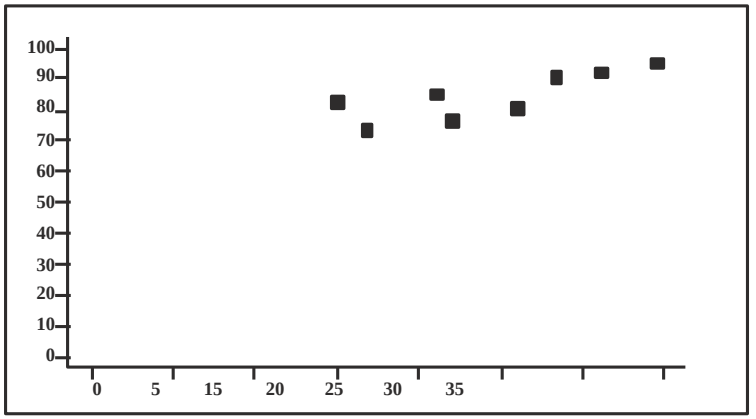
Example:

The grade of MTH 112 examination at the end of rain semester of 2006/2007 session for 8 students given the hours of studied out side class work are given below:

Student	1	2	3	4	5	6	7	8
Hours of study (x)	20	13	10	23	27	32	18	22
Exam Score (y)	64	61	80	75	70	90	72	70

Plot a scatter diagram for the data and obtain the regression equation.

SCATTER DIAGRAM



Mean of $X = \frac{\sum x}{N} = \frac{165}{8} = 20.625$

Mean of $Y = \frac{\sum xy}{N} = \frac{582}{8} = 72.75$

The regression line is given as

Y = a + bx

Semi-average

This method consists of splitting the data into two equal groups. Plotting the mean points for each group these two points with a straight line.

Example:

Using the example above

Student	1	2	3	4	5	6	7	8
Hours of study (x)	20	13	10	23	27	32	18	22
Exam Score (y)	64	61	80	75	70	90	72	70
	10	13	18	20	22	23	27	32
	80	61	72	64	70	75	70	90

Total 1st group	*Second group*
X=61	104
Y = 277	305
Mean 1st group	
X = 15.25	26
Y = 69.25	76.25

The regression equation y = a + bx can be determined by locating y intercept and the slope of the line.

THE LEAST SQUARES METHOD

The least squares regression line of y on x is that line for which the sum of squares of the vertical deviation of all points from the line is least, i.e. the line of best fit is one in which, $\sum d^2$ for all i is least.

In other words the sum of the squared deviations of the observed data point (y) from the least squares line is smaller than the sum of the squared deviations of the data point from any other line that can be drawn through the data point. It is for this reason that the line is called the least squares line.

SIMPLE LINEAR REGRESSION

This is when there is only one independent variable x with a single dependent variable (y) of which it is desirable to determine the relationship between x and y. The independent variable x is usually assumed to be a continuous variable which is under the control of the experiment. Then if the experiment is designed we choose the valve of x and observe the corresponding value of y.

$$E(y) = \beta 0 + \beta i x$$

Where the parameter of the straight line ß and $ß_1$ are unknown constants. Each observation y can be described by

a model

$Y_i = \beta_o + \beta_i X_i + C_i$ such that $i = 1, 2, \ldots, n$ and the least squares function is $L = y_i - \beta_o - \beta_i x_i - C$

$$L = \sum_{i=1}^{n} \sum^2 = \sum_{i=1} [y_i - \beta_o - \beta_i X_i]^2$$

$$\frac{\partial L}{\partial \beta_i} = 2^n \sum_{i=1} X_i [y_i - \beta_o - \beta_i X_i]$$

- (1)

$$\frac{\partial L}{\partial \beta_o} = 2 \sum_{i=1}^{3} [y_i - \beta_o - \beta_i X_i]$$

$$\partial L = 0 \Rightarrow -2^n \sum (y_i - \beta_o - \beta_i X_i) = 0$$

Divides by -2

$$\sum X_i Y_i - \sum X_i \beta_o - \beta_i \sum X_i^2 = 0$$

- (2)

$$\frac{\partial L}{\partial \beta_i} = 2^n \sum X_i (y_i - \beta_o - \beta_i X_i)$$

$$\partial L = 20 \Rightarrow -2 \sum (y_i - \beta_o - \beta_i X_i) = 0$$

Divide by -2 we have

$$\sum Y_i - n\beta_o - \beta_i \sum X_i = 0$$

$$n\beta_o = \sum Y_i - \beta_i \sum X_i = 0$$

$$\beta_o = \frac{\sum Y_i}{n} - \frac{\beta_i \sum X_i}{n} = 0$$

$$\beta = Y - \beta_i X$$

From 2, $\sum X_i Y_i - \sum X_i \beta_o - \beta_i \sum X_i^2$

$$\sum X_i Y_i - \sum X_i \left(\frac{\sum Y_i}{n} - \frac{\beta_i \sum X_i}{n}\right) - \beta_i \sum X_i^2 = 0$$

$$\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n} + \frac{\beta_i (\sum X_i)^2}{n} - \beta_i \sum X_i^2$$

$$\beta_i \sum X_i^2 - \beta_i \frac{(\sum X_i)^2}{n} = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$$

$$\beta_i = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - \sum X_i^2}$$

Example

The grade of MTH 212 examination at the end of the first semester 2006/2007 session for 8 students given the hours of study outside class work are;

Student	1	2	3	4	5	6	7	8
Hours of study (x)	20	13	10	23	27	32	18	22
Exam Score (y)	64	61	80	75	70	90	72	70

Fit an appropriate model to this data

Solution

$Y = ß_1 + ß_1 X_1$

$ß_1 = Y - ß_1 X$

$ß_1 = \frac{\Sigma xy - \Sigma xy}{\Sigma x^2 - (\Sigma x)^2}$

$\Sigma XY = 20.625$ $\quad \Sigma\chi^2 = 3759$ $\quad \Sigma = 165$

$X = 20.625$ $\quad \Sigma Y = 72.75$ $\quad \Sigma Y = 582$

$ß_1 = \frac{8 \times 3759 (165)}{8x3759 - (165)^2.}$

$\frac{1602}{2840} = 0.564$

$ß_1 = 0.564$

-

$ß_o = Y - ß_1 = 72.25 - 0.564 \times 20.625$

$B_o = 61.118$

Y=61.118 +0.564x

582x165.

USES OF REGRESSION EQUATION

One of the uses of regression equation is for prediction. Prediction can be adopted when,

(i) It is possible to observe an x value and impossible or impractical to observe a corresponding y value.

(ii) One wishes to say something about a particular value of y

(iii) It is used for comparison purposes.

References

- United States Census Bureau. (2010). Post Enumeration Survey.
- National Population Commission of Nigeria. (2006). Report on Post Enumeration Survey.
- Moore, D. S., McCabe, G. P., & Craig, B. A. (2018). Introduction to the Practice of Statistics. W.H. Freeman and Company.

Triola, M. F. (2018). Elementary Statistics. Pearson