PRIMARY AND SECONDARY DATA

Primary Data: Primary data refers to firsthand information collected directly from original sources for a specific purpose. This data is collected through surveys, interviews, experiments, observations, and questionnaires. For example, a researcher conducting a survey to understand students' study habits, a company gathering customer feedback through interviews.

Merits of Primary Data:

1. Primary data are directly related to the research objective.
2. Primary data are collected in real-time and therefore, it ensures accuracy.
3. Primary data provides greater control to the research objective. Researcher decides the methodology and data collection process.
4. Primary data is more authentic and it is free from manipulation, which make it highly reliable.

Demerits of Primary Data:

1. Primary data collection is more time consuming. Also, data collection requires significant effort and planning.
2. Primary data is costly as surveys, interviews, and experiments can be expensive.
3. Primary data has limited scope. As data is collected from specific samples, therefore, it is less generalizable.

Secondary Data

Secondary data is information that has already been collected, analyzed, and published by someone else. This data comes from books, research papers, government reports, newspapers, and online sources. For example, using census data for population analysis.

Merits of Secondary Data:

1. Secondary data is easily available.
2. Secondary data is also cost-effective.
3. Secondary data saves time and efforts.
4. Secondary data covers large populations over long periods.

Demerits of Secondary Data:

1. Secondary data may be outdated and therefore it may not reflect current conditions.
2. Secondary data may not fit the exact research needs.

3. The reliability of secondary data depends on the source's credibility.

4. In secondary data researchers have no control over how data was collected.

When to Use Primary and Secondary Data?

- **Primary Data** is best used when conducting **original research** for specific objectives, such as market surveys, scientific experiments, and field studies.
- **Secondary Data** is useful when gathering **background information**, comparing trends, or conducting a **preliminary analysis** before collecting primary data.

Differences Between Primary and Secondary Data

The differences between secondary and primary data are as follows:

1) Primary data is collected firsthand, while secondary data is collected by others.
2) Primary data specific to research needs while secondary data specific to general purpose.
3) Primary data is more expensive as compared to secondary data.
4) Primary data is time consuming, while secondary data is quick to obtain.
5) The accuracy and reliability of primary data is high. But secondary data may contain errors or bias.
6) In primary data, researchers control data collection. But in secondary data researcher have no control over data collection.

POPULATION AND SAMPLE

Population: A **population** refers to the entire group of individuals, items, or data points that share a common characteristic and are the focus of a study. It includes every possible observation relevant to the research. For example, all college students in India.

Sample: A **sample** is a smaller, selected group from the population that is studied to make inferences about the entire population. Since studying an entire population is often impractical, researchers analyze a sample to save time and resources. For example, 500 students randomly chosen from different universities in India.

Differences between Population and Sample: The followings are the differences between population and sample:

1) A population includes all individuals or items in a study, whereas a sample is a selected subset of the population.
2) The population is usually large or infinite, while a sample is smaller and manageable.
3) Studying the entire population is difficult and time-consuming, whereas a sample can be studied efficiently.
4) A population provides complete and exact data, while a sample gives an estimate of the population's characteristics.

5) Collecting data from a population requires more time and resources, whereas a sample-based study is quicker and cost-effective.

Population Parameter and Sample Statistic

Population Parameter: A **population parameter** is a numerical value that describes a characteristic of an entire population. Since populations are large, parameters are often unknown and need to be estimated. For example, the average height of all students in a country.

Sample Statistic: A **sample statistic** is a numerical value calculated from a sample, used to estimate a population parameter. Since studying the entire population is difficult, researchers use statistics from samples to make inferences. For example, the average height of 500 randomly selected students from different schools.

Differences Between Population Parameter and Sample Statistic

1. A population parameter describes a characteristic of an entire population, while a sample statistic describes a characteristic of a selected sample.
2. Parameters are denoted by Greek letters (e.g., μ for mean, σ for standard deviation), while statistics are represented by Latin letters (e.g., $\bar{x}$ for mean, s for standard deviation).
3. Population parameters are usually unknown and need to be estimated, whereas sample statistics are calculated directly from collected data.
4. A parameter is based on the whole population, making it more accurate, while a statistic is based on a sample and may have some level of error.
5. A parameter gives a true value for the entire population, while a statistic helps estimate the parameter when studying the whole population is impractical.

Principal Steps in a Sample Survey

A sample survey involves systematically collecting data from a selected group (sample) to infer characteristics about a larger population. The key steps in conducting a sample survey are:

1. Defining the Objective: The first step in a sample survey is to define the objective of the study. The researcher should clearly outline the purpose of the survey and identify the key variables to be measured.

2. Defining the Population: The target population should be determined and from which the sample will be drawn.

3. Choosing the Sampling Method: To select sample, a researcher should decide on a suitable sampling technique such as random sampling, stratified sampling, cluster sampling, systematic sampling, etc.

4. Determining the Sample Size: Sample size determination is an important step in the process of sample survey. Select an appropriate sample size to ensure accuracy while keeping the survey manageable. A larger sample improves reliability but increases costs.

5. Designing Data Collection Method: Suitable data collection method is one of the prime concerns of sample survey.

6. Data Collection: After preparing data collection methods, gather information from the selected sample using face-to-face interviews, online surveys, phone calls, or mailed questionnaires.

7. Data Processing: The collected data should be checked for errors, inconsistencies, and missing values. After this step it is essential to organize data for analysis through coding and classification.

8. Data Analysis and Interpretation: For data analysis use statistical methods to summarize and interpret the results. One may also identify patterns, trends, and relationships in the data.

9. Drawing Conclusions and Making Inferences: Generalize findings from the sample to the entire population. Acknowledge limitations and possible sources of error.

10. Preparation and Presentation of Report: Present findings in a clear and structured report. Use graphs, charts, and tables to enhance understanding.

11. Making Policy Recommendations: Based on survey findings, suggest actions or strategies for decision-makers.

DIAGRAMMATIC REPRESENTATION OF STATISTICAL DATA

Diagrammatic representation is a visual method of presenting statistical data that makes complex information easier to understand and interpret. There are several types of diagrammatic representations, each suited for different types of data and analysis.

1. Bar Diagram

Method of Construction:

1. Draw two axes:
    - The **X-axis** represents categories (e.g., cities, years, industries).
    - The **Y-axis** represents numerical values (e.g., population, sales, profits).
2. Choose an appropriate scale for the Y-axis.
3. Draw bars of equal width with gaps between them.

4. Label the bars clearly and give a title to the diagram.

Uses:

- Used for **comparisons** between different categories.
- Applied in **population studies, business statistics, and social research**.

Merits:

a. Simple to construct and easy to interpret.
b. It can be used for both small and large datasets.
c. Useful for making quick comparisons.

Demerits:

a. Limited use for continuous data.
b. Difficult to use for a large number of categories.

Types of Bar Diagrams:

- **Simple Bar Diagram**: Represents a single variable (e.g., annual rainfall in different cities).
- **Multiple Bar Diagram**: Compares two or more variables side by side (e.g., male vs. female literacy rates).
- **Stacked (Component) Bar Diagram**: Shows the breakdown of subcategories within a bar (e.g., total expenditure divided into rent, food, and education).

2. Pie Chart

Method of Construction:

1. Convert each category's value into a percentage of the total.
2. Multiply each percentage by **360°** to get the sector angle.
3. Draw a circle and use a protractor to mark each sector accordingly.
4. Label each sector with the category name and percentage.

Uses:

- Best for showing **percentage distribution** (e.g., market share, budget allocation).
- Commonly used in **business, economics, and demography**.

Merits:

a) Gives a clear visual comparison of proportions.
b) Effective for data with a limited number of categories.

Demerits:

a) Not suitable for large datasets with many categories.
b) Difficult to compare slight variations between sectors.

3. Histogram

Method of Construction:

1. Draw two axes:
    - **X-axis**: Represents class intervals (continuous data).
    - **Y-axis**: Represents frequency.
2. Choose a scale and draw adjacent bars (unlike bar diagrams, there are no gaps).
3. Height of each bar represents the frequency of that class interval.

Uses:

- Used for **continuous data representation**, such as income distribution, age groups, and temperature variations.
- Helps in **identifying patterns, such as normal distribution**.

Merits:

a) Ideal for showing frequency distribution.
b) Useful for identifying skewness and distribution shape.

Demerits:

a) It cannot be used for discrete or categorical data.
b) Exact values are difficult to extract from the graph.

4. Frequency Polygon

Method of Construction:

1. Construct a histogram.
2. Identify the midpoints of each class interval.
3. Plot points corresponding to the midpoints and their frequencies.
4. Connect the points with straight lines.
5. Extend the polygon to the X-axis for closure.

Uses:

- Used for **continuous data** where trends and distributions need to be analyzed.
- Common in **statistical research, economics, and science experiments**.

Merits:

a) Useful for comparing multiple distributions.
b) Smoother than histograms, making trends clearer.

Demerits:

a) Requires accurate midpoint calculations.
b) May be confusing if too many distributions are plotted.

5. Ogive (Cumulative Frequency Curve)

Method of Construction:

1. Construct a cumulative frequency table.
2. Plot points corresponding to cumulative frequencies and class intervals.
3. Join the points with a smooth curve.

Uses:

- Used for **finding median, quartiles, and percentiles**.
- Helps analyze **cumulative trends** (e.g., income levels in a population).

Merits:

a) Useful for percentile and median estimation.
b) Easy to interpret cumulative trends.

Demerits:

a) Less effective for comparing multiple datasets.
b) Cannot be used for discrete data.

6. Line Graph

Method of Construction:

1. Plot values on the X and Y axes.
2. Connect the points using straight lines.
3. Label axes and provide a title.

Uses:

- Used for **time series data** (e.g., stock prices, sales growth, temperature changes).
- Helps visualize **trends and fluctuations**.

Merits:

a) Effective for showing trends over time.
b) Easy to interpret changes and patterns.

Demerits:

a) Not suitable for large datasets with too much variation.
b) May become misleading if scales are not chosen properly.

7. Scatter Diagram

Method of Construction:

1. Plot two variables on the X and Y axes.
2. Mark data points for each pair of values.
3. Observe the pattern of points to determine correlation.

Uses:

- Used for **analyzing relationships** between two variables (e.g., income vs. expenditure, height vs. weight).
- Common in **correlation and regression analysis**.

Merits:

a) Best for identifying relationships between variables.
b) Provides a basis for correlation analysis.

Demerits:

a) Does not show exact numerical relationships.
b) Requires further statistical methods to confirm correlations.

8. Pictograms

Method of Construction:

1. Select symbols or images representing the data.
2. Ensure that each symbol represents a fixed value (e.g., 1 icon = 100 people).

3. Arrange the symbols proportionally to represent the data.

Uses:

- Used for **public communication, newspapers, and social media**.
- Ideal for representing **data in an engaging way**.

Merits:

a) Easy to understand, even for non-statistical audiences.
b) Visually appealing and engaging.

Demerits:

a) Can be misleading if proportions are not accurate.
b) Limited use for precise or complex statistical data.

Merits of Diagrammatic Representation

1. **Easy to Understand**: Visual representation makes complex data more accessible.
2. **Quick Comparison**: It helps compare trends, proportions, and relationships at a glance.
3. **Identifies Trends and Patterns**: It is useful for spotting upward/downward trends in time series data.
4. **Saves Time**: It requires less effort to interpret compared to numerical tables.
5. **Effective Communication**: It is useful for presentations, reports, and decision-making.
6. **Highlights Key Insights**: Makes it easier to focus on important findings.
7. **Attractive and Engaging**: More appealing than raw data tables, especially for the general audience.

FREQUENCY DISTRIBUTION

A **frequency distribution** is a way of organizing raw data into a structured table that shows the number of times (frequency) each value or group of values occurs in a dataset. It helps in summarizing large datasets and identifying patterns.

Types of Frequency Distribution

1. **Discrete Frequency Distribution**: It is used for data with distinct values (e.g., number of students in a class).

2. **Continuous Frequency Distribution**: It is used for data grouped into class intervals (e.g., height ranges of students).
3. **Cumulative Frequency Distribution**: It shows the cumulative total of frequencies up to each class interval.

Steps to Construct a Frequency Table

1. Collect the Data: Gather raw data that needs to be organized (e.g., marks of 50 students in a test).

2. Determine the Range: Find the difference between the highest and lowest values in the dataset.

3. Decide the Number of Classes: The number of class intervals (bins) should be chosen appropriately (usually **5 to 10**).

4. Determine Class Intervals: Choose a suitable class width (interval size) by dividing the range by the number of classes.

5. Tally the Data: Count how many values fall into each class interval. Use tally marks to keep track of occurrences.

6. Record Frequencies: Write the total count for each class in the frequency column.

METHODS OF SAMPLING

Introduction

Sampling is the process of selecting a sample from a larger population to analyze and draw conclusions. Since studying an entire population is often impractical due to time, cost, and effort constraints, sampling helps obtain reliable results with minimal resources.

There are two broad categories of sampling methods:

1. **Probability Sampling (Random Sampling)**
2. **Non-Probability Sampling**

1. Probability Sampling (Random Sampling): Probability sampling ensures that every unit in the population has a known and equal chance of being selected. This method provides more representative and unbiased results, making it ideal for statistical analysis.

(i) Simple Random Sampling: Each unit in the population is chosen randomly using methods like a lottery system or random number tables.

Steps to Conduct:

1. Assign a unique number to each member of the population.
2. Use a random selection method (lottery, random number generator, etc.).
3. Select the required number of samples.

Merits:

- Eliminates selection bias.
- Simple and easy to implement for small populations.
- Ensures equal chances of selection.

Demerits:

- Not feasible for large populations.
- Requires a complete list of the population.

(ii) Systematic Sampling: Every k-th item from a population is selected after choosing a random starting point.

Steps to Conduct:

1. Determine the total population size (N) and required sample size (n).
2. Calculate the sampling interval: **k = N/n**.
3. Choose a random starting point between 1 and k.
4. Select every k-th unit from the population.

Merits:

- Easier and quicker than simple random sampling.
- Ensures even distribution across the population.

Demerits:

- May introduce bias if there is a hidden pattern in the population.
- Less random than simple random sampling.

(iii) Stratified Sampling: The population is divided into homogeneous groups (strata) based on characteristics such as age, income, or education, and random samples are taken from each group.

Steps to Conduct:

1. Identify relevant strata based on characteristics.
2. Determine the proportion of the population in each stratum.
3. Use random sampling within each stratum to select participants.

Merits:

- More accurate and representative.
- Ensures all groups are included in the sample.

Demerits:

- Requires prior knowledge of population characteristics.
- Complex and time-consuming.

(iv) Cluster Sampling: The population is divided into clusters (e.g., schools, cities), and entire clusters are randomly selected instead of individuals.

Steps to Conduct:

1. Divide the population into clusters.
2. Select clusters randomly.
3. Study all elements within the selected clusters.

Merits:

- Cost-effective and time-saving for large populations.
- Useful when population data is unavailable.

Demerits:

- Can introduce bias if selected clusters are not representative.
- Less precise compared to other probability methods.

2. Non-Probability Sampling

Non-probability sampling does not provide every unit in the population with a known chance of selection. This method is useful for exploratory research but may introduce bias.

(i) Convenience Sampling: Selection is based on ease of access (e.g., surveying people at a shopping mall).

Merits:

- Quick and inexpensive.
- Suitable for preliminary research.

Demerits:

- High selection bias.
- Findings may not be generalizable.

(ii) Judgmental (Purposive) Sampling: The researcher selects participants based on personal judgment of who will provide the most useful information.

Merits:

- Useful for expert opinions and specialized studies.
- Saves time when targeting specific groups.

Demerits:

- Highly subjective and prone to bias.
- Results may not be representative.

(iii) Quota Sampling: The researcher selects a sample based on predetermined quotas (e.g., 50% men and 50% women).

Merits:

- Ensures representation of key groups.
- Faster and easier than stratified sampling.

Demerits:

- Selection bias due to non-randomization.
- Less accurate than probability sampling.

(iv) Snowball Sampling: Existing participants refer other participants, creating a **network effect** (e.g., studying hard-to-reach populations like drug users or refugees).

Merits:

- Useful for studying hidden or rare populations.
- Cost-effective when population lists are unavailable.

Demerits:

- Prone to bias as referrals may be from similar backgrounds.
- Difficult to generalize findings.

Differences Between Random and Non-Random Sampling: The following are the differences between random sampling and Non-Random Sampling:

1) A sampling method where every unit in the population has an equal chance of being selected is called random sampling. But in non-random sampling method of selection of samples is based on subjective judgment or convenience, without equal chances for all units.

2) Basis of Selection: In random sampling, selection is done through random methods like lottery, random number generators, or systematic rules. But in non-random sampling, selection is based on researcher preference, accessibility, or predefined criteria.

3) Random sampling is less biased since every unit has an equal chance of selection. But non-random sampling is prone to bias as the researcher influences the selection process.

4) Random Sampling produces a more representative sample, making it suitable for generalizing results for the population. But non-random sampling may not accurately represent the entire population, leading to limited generalizability.

5) Random sampling includes Simple Random Sampling, Systematic Sampling, Stratified Sampling, and Cluster Sampling. But non-random sampling includes Convenience Sampling, Judgmental Sampling, Quota Sampling, and Snowball Sampling.

6) Random sampling is best for large populations and when statistical accuracy is needed, while non-random sampling is best for exploratory research, case studies, or when population data is unavailable.

7) Random sampling is more complex and expensive due to the need for proper planning and data collection methods. But non-random sampling is easier, quicker, and less costly as it does not require a complete list of the population.

8) Random sampling is commonly used in scientific, medical, and social science research to ensure unbiased results. But non-random sampling used in qualitative research, pilot studies, and exploratory studies where statistical precision is not the primary focus.