Eco Industry
Oriented University

sinarmas

# SAMPLING AND SURVEY TECHNIQUES

Study Case Using R and Python


**Writer:**

Bakti Siregar, M.Sc., CDS.


D'SCIENCELABS
Smart Idea

Kampus Merdeka
INDONESIA JAYA


First Edition

# Sampling and Survey Techniques

Study Case Using R and Python

Bakti Siregar, M.Sc.,CDS

# Table of contents

In today's data-driven world, the ability to collect, analyze, and interpret information accurately is more important than ever. Sampling and survey techniques are essential tools in research, business, social sciences, and public policy, allowing researchers to gather representative data, identify patterns, and make well-informed decisions. A well-structured survey, combined with appropriate sampling methods, enhances data reliability while minimizing bias and errors.

The foundation of effective surveys lies in selecting the right sampling strategy. Probability-based sampling methods, such as simple random sampling, stratified sampling, and cluster sampling, provide statistically valid insights, whereas non-probability techniques, including convenience sampling and quota sampling, offer practical advantages in specific research contexts. Understanding these methods allows researchers to optimize data collection while ensuring the accuracy and credibility of their findings.

Beyond data collection, survey methodology involves data validation, analysis, and interpretation. Statistical techniques such as confidence intervals, regression analysis, and hypothesis testing enhance the quality of survey results, allowing researchers to derive meaningful insights. Additionally, modern digital tools and automation have transformed survey research, improving efficiency and accessibility.

This book provides a comprehensive and practical guide to sampling and survey techniques, covering the fundamental principles of data collection, survey design, statistical analysis, and visualization. By mastering these concepts, researchers, analysts, and decision-makers will be better equipped to design effective surveys, interpret survey data accurately, and apply the results to real-world decision-making processes.

# Preface

## About the Writer

Bakti Siregar, M.Sc., CDS works as a Lecturer at the ITSB Data Science Program. He earned his Master's degree from the Department of Applied Mathematics at National Sun Yat Sen University, Taiwan. In addition to teaching, Bakti also works as a Freelance Data Scientist for leading companies such as JNE, Samora Group, Pertamina, and PT. Green City Traffic.

He has a strong enthusiasm for projects (and teaching) in the fields of Big Data Analytics, Machine Learning, Optimization, and Time Series Analysis, particularly in finance and investment. His core expertise lies in statistical programming languages such as R Studio and Python. He is also experienced in implementing database systems like MySQL/NoSQL for data management and is proficient in using Big Data tools such as Spark and Hadoop. Some of his projects can be viewed here: Rpubs, Github, Website, and Kaggle.

---

## Acknowledgments

**Sampling and Survey Techniques** play a crucial role in data collection, ensuring accuracy and representativeness in statistical analysis. This book provides fundamental concepts and practical applications of various sampling methods and survey designs used in research and industry.

I appreciate the active participation of students, researchers, and practitioners, whose questions and discussions enriched the learning experience. I hope this book serves as a practical guide for applying proper sampling techniques in real-world scenarios.

## Feedback & Suggestions

Your feedback is essential in improving this book. We invite all readers to share their thoughts on the content, structure, and clarity of the materials. Suggestions for additional topics or areas requiring further explanation are highly appreciated.

With your support and contributions, we aim to refine this book, making it a more comprehensive resource for **Sampling and Survey Techniques**. Thank you for your participation!

For feedback and suggestions, feel free to contact:


- dsciencelabs@outlook.com
- siregarbakti@gmail.com
- siregarbakti@itsb.ac.id

# Chapter 1

# Principles of Sampling

## 1.1 What is Sampling?

Sampling is the process of selecting a subset of individuals, items, or observations from a larger population to estimate characteristics of the whole population. It is widely used in research, business, and public policy to make data-driven decisions efficiently.

## 1.2 Population vs. Sample

In statistics, understanding the distinction between **population** and **sample** is crucial for data analysis, inference, and decision-making.

### 1.2.1 Population

The **population** ($N$) is the **entire group** of individuals, objects, or events that a researcher is interested in studying. It includes **all possible observations** relevant to the research.

- **Examples:**
  - All residents of a city when studying voting behavior.
  - Every manufactured smartphone from a factory when analyzing defect rates.
  - Every student in a university when measuring average exam scores.
- **Types of Populations:**
  - **Finite Population:** A population with a fixed number of elements (e.g., employees in a company).
  - **Infinite Population:** A population with an uncountable number of elements (e.g., bacteria in a petri dish).
  - **Target Population:** The specific population a researcher wants to study.
  - **Accessible Population:** The portion of the target population available for study.

### 1.2.2 Sample

A **sample** ($n$) is a **subset** of the population, selected for analysis. Since studying an entire population is often impractical due to cost, time, or accessibility, a sample is used to make inferences about the population.

- **Examples:**
  - Surveying 1,000 residents of a city to estimate public opinion.
  - Inspecting 500 randomly chosen smartphones to assess defect rates.
  - Analyzing exam scores from 200 randomly selected students.
- **Characteristics of a Good Sample:**
  - **Representative:** Accurately reflects the population.
  - **Random:** Selected without bias.
  - **Sufficiently Large:** Ensures reliable estimates.
  - **Minimally Biased:** Avoids systematic errors.

---

### 1.2.3 Key Differences

When conducting research or statistical analysis, it is essential to distinguish between population and sample. The population refers to the entire group of interest in a study, while the sample is a smaller subset selected from that population for analysis. Understanding their differences is crucial for making accurate inferences and ensuring the validity of conclusions.

Here are the key differences between a population and a sample:

| Feature | Population ($N$) | Sample ($n$) |
|---|---|---|
| **Definition** | Entire group of interest | A subset selected for study |
| **Size** | Large or infinite | Smaller, manageable portion |
| **Notation** | Uses uppercase letters (e.g., $N$, $\mu$, $\sigma$) | Uses lowercase letters (e.g., $n$, $\bar{x}$, $s$) |
| **Parameters** | True values (e.g., population mean $\mu$, standard deviation $\sigma$) | Estimates (e.g., sample mean $\bar{x}$, standard deviation $s$) |
| **Cost & Time** | High | Lower |
| **Accuracy** | Provides exact information | Provides an estimate with some margin of error |

---

## 1.3 Why Use a Sample?

In research and data collection, studying an entire population is often impractical or impossible. Instead, researchers use a **sample**, which is a smaller, manageable subset of the population. Below are the key reasons for using a sample:

1. **Cost-Effectiveness**
   Collecting data from an entire population requires significant financial resources. A sample reduces costs associated with data collection, processing, and analysis.

2. **Time Efficiency**
   Studying an entire population is time-consuming. A well-chosen sample allows for quicker data collection and analysis.

3. **Feasibility**
   Some populations are too large or inaccessible to study completely. A sample makes research possible when population-wide data collection is impractical.

4. **Accuracy and Reliability**
   When selected properly, a sample can provide highly accurate and reliable insights. Statistical techniques ensure that the sample represents the entire population effectively.

5. **Reduced Data Management Complexity**
   Handling vast amounts of data can be challenging. A sample simplifies data management while still providing meaningful conclusions.

6. **Ethical Considerations**
   Some research (e.g., medical trials) may involve risks, making it unethical to test on an entire population. A sample allows for controlled and ethical experimentation.

---

## 1.4 Avoiding Sampling Bias

**Sampling bias** occurs when certain members of the population are systematically **excluded** or **overrepresented** in the sample. This leads to inaccurate and unrepresentative results, potentially skewing conclusions and reducing the validity of a study.There are some sauses of sampling bias:

- **Undercoverage:** Some groups in the population are not included in the sampling frame.

- **Overrepresentation:** Certain groups have a disproportionately higher chance of being selected.

- **Self-Selection Bias:** Participants voluntarily choose to be part of the study, leading to a non-random sample.

Therefore, consider the following to minimize sampling bias:

1. **Ensure a Representative Sampling Frame**
   Use an up-to-date and comprehensive list that includes all relevant members of the population. Avoid missing subgroups that may be critical to the study's findings.

2. **Use Random Selection Techniques**
   Implement **simple random sampling**, **stratified sampling**, or **systematic sampling** to ensure each member of the population has an equal chance of selec-

tion. Avoid **convenience sampling**, which selects participants based on ease of access rather than representativeness.

3. **Avoid Self-Selection Bias**
   Prevent situations where individuals **opt-in** to participate in a way that skews results. Use **randomized invitations** rather than voluntary participation. Consider offering incentives to encourage a broader and more diverse group of respondents.

Minimizing sampling bias is essential for producing valid, reliable, and generalizable research findings. By ensuring a well-constructed sampling frame, applying random selection methods, and reducing self-selection effects, researchers can improve the quality and accuracy of their studies.

## 1.5 Randomization in Sampling

Randomization is a process that ensures every member of a population has an **equal chance** of being selected. This reduces **sampling bias** and enhances the **generalizability** of research findings.

### 1.5.1 Simple Random Sampling

A method where each element in the population has an equal probability of selection, ensuring a truly random sample. Here, how it works:

- Assign a unique number to each member of the population.
- Use a random number generator or lottery system to select participants.

**Example:** A company wants to survey 500 employees from a workforce of 5,000. Each employee is assigned a number, and 500 are randomly chosen using a lottery system.

### 1.5.2 Systematic Sampling

A method where elements are selected at regular intervals from an ordered list. Here, how it works:

- Determine the sample size (e.g., selecting 100 people from a list of 1,000).
- Calculate the sampling interval: **Population Size ÷ Sample Size** (e.g., 1,000 ÷ 100 = 10).
- Randomly select a starting point and then pick every 10th person.

**Example:** A researcher wants to survey every 5th customer from a list of 1,000 shoppers. If the starting point is 3, the selected individuals will be 3rd, 8th, 13th, etc.

### 1.5.3 Stratified Sampling

A method that divides the population into **subgroups (strata)** based on a shared characteristic, then randomly selects a proportional number of participants from each stratum. Here, how it works: - Identify relevant **strata** (e.g., age groups, income levels,

education).
- Determine the proportion of each stratum in the population.
- Conduct **random sampling** within each stratum.

**Example:** A university wants to survey students from different academic years. If 40% of students are freshmen, 30% are sophomores, 20% are juniors, and 10% are seniors, then the sample will reflect these proportions.

Using random sampling methods like **SRS, systematic sampling, and stratified sampling** helps ensure a **fair, unbiased**, and **representative sample**. This improves the reliability and validity of research findings, making them more generalizable to the entire population.

## 1.6 Challenges in Sampling

Sampling is a critical process in research, but it comes with several challenges that can impact accuracy and reliability. Below is an overview of key sampling challenges along with their causes and possible solutions.

| Challenge | Causes | Solutions |
|---|---|---|
| Non-Response Bias | Participants unwilling or unable to respond. Surveys too long or complex. Certain groups less likely to participate. | Send follow-up reminders. Offer incentives. Simplify survey format. |
| Sampling Frame Errors | Outdated or incomplete lists. Incorrect classification. Duplicate or ineligible participants included. | Keep the sampling frame updated. Cross-check data sources. Use stratified sampling. |
| Inadequate Sample Size | Limited resources for large samples. Miscalculated sample size. High dropout rates in longitudinal studies. | Use statistical methods to determine the correct sample size. Account for potential dropouts. |
| Cost and Time Constraints | High costs for data collection. Delays in reaching participants. Need for specialized tools or personnel. | Use cost-effective methods like online surveys. Automate data collection. Optimize resources. |

Addressing these challenges ensures that the sampling process is more reliable, efficient, and representative of the target population. By implementing effective solutions, researchers can minimize errors and improve the overall quality of their studies.

## 1.7 Applications in Industry

Sampling plays a crucial role across various industries, allowing organizations to gather insights, make informed decisions, and optimize processes. Below are key areas where sampling is widely used:

| Industry | Application | Purpose |
|---|---|---|
| Market Research | Conducting surveys and focus groups. | Understanding customer preferences, trends, and behaviors. |
| Healthcare | Studying patient data and clinical trials. | Estimating disease prevalence, treatment effectiveness, and public health trends. |
| Quality Control | Inspecting a subset of products in manufacturing. | Ensuring product quality and compliance with industry standards. |
| Finance | Analyzing financial transactions and market trends. | Assessing risks, detecting fraud, and making investment decisions. |

By applying proper sampling techniques, industries can obtain **accurate and reliable insights** while minimizing errors and biases. This ensures better decision-making, cost savings, and improved operational efficiency.

# Chapter 2

# Probability and Distributions

## 2.1 Probability in Sampling

## 2.2 Probability vs. Non-Probability Sampling

## 2.3 Types of sampling distributions

### 2.3.1 Sampling distribution of mean

### 2.3.2 Sampling distribution of proportion

### 2.3.3 T-distribution

## 2.4 Central Limit Theorem

## 2.5 Law of Large Numbers

## 2.6 Confidence Intervals

## 2.7 Hypothesis Testing in Surveys

# Chapter 3

# Sampling Methods

## 3.1 Overview of Sampling Methods

## 3.2 Probability Sampling

### 3.2.1 Simple Random Sampling

### 3.2.2 Stratified Sampling

### 3.2.3 Cluster Sampling

### 3.2.4 Systematic Sampling

## 3.3 Non-Probability Sampling

### 3.3.1 Convenience Sampling

### 3.3.2 Quota Sampling

### 3.3.3 Snowball Sampling

### 3.3.4 Judgmental Sampling

## 3.4 Choosing the Right Method

## 3.5 Strengths & Limitations

## 3.6 Real-World Applications

# Chapter 4

# Sample Size and Margin of Error

# Chapter 5

# Questionnaire Design

# Chapter 6

# Digital Data Collection

**6.1 Introduction to Digital Data Collection**

**6.2 Advantages of Digital Surveys**

**6.3 Online Survey Platforms**

**6.4 Mobile Data Collection**

**6.5 Web-Based vs. App-Based Surveys**

**6.6 Automated Data Entry & Processing**

**6.7 Ensuring Data Security & Privacy**

**6.8 Reducing Errors in Digital Surveys**

**6.9 Real-Time Data Monitoring**

**6.10 Challenges & Best Practices**

# Chapter 7

# Data Validation

## 7.1 Introduction to Data Validation

## 7.2 Importance of Data Quality

## 7.3 Common Data Errors in Surveys

## 7.4 Techniques for Data Cleaning

## 7.5 Automated vs. Manual Validation

## 7.6 Handling Missing Data

## 7.7 Detecting Outliers & Inconsistencies

## 7.8 Duplicate Response Detection

## 7.9 Validation in Digital Surveys

## 7.10 Best Practices for Data Accuracy

# Chapter 8

# Seven Tools Analysis

## 8.1 Introduction to the Seven Tools

### 8.1.1 Check Sheets for Data Collection

### 8.1.2 Histograms for Data Distribution

### 8.1.3 Pareto Charts for Prioritization

### 8.1.4 Cause-and-Effect (Ishikawa) Diagrams

### 8.1.5 Scatter Plots for Relationship Analysis

### 8.1.6 Control Charts for Process Monitoring

### 8.1.7 Flowcharts for Process Visualization

## 8.2 Applying the Tools in Surveys & Sampling

## 8.3 Real-World Case Studies

# Chapter 9

# Visualization and Reporting

## 9.1 Importance of Visualization

## 9.2 Summarizing Survey Results

### 9.2.1 Check Sheets

### 9.2.2 Bar Charts

### 9.2.3 Pie Charts

### 9.2.4 Histograms

### 9.2.5 Distributions

### 9.2.6 Scatter Plots

### 9.2.7 Pareto Charts

## 9.3 Survey Dashboards

## 9.4 Automated Reporting

## 9.5 Data Interpretation

## 9.6 Best Practices

# Chapter 10

# Surveys in Risk Analysis

# Chapter 11

# Decision Making with Surveys

# Chapter 12

# Case Studies