# INTRODUCTION TO STATISTICS

## A Data Science Perspective
### with

**R**



**Written by:**

Bakti Siregar, M.Sc., CDS.

# Introduction to Statistics
## A Data Science Perspective with R

Bakti Siregar, M.Sc., CDS.

# Table of contents

We live in a world overflowing with data. From science and business to policy and everyday life, the ability to interpret data through statistics has become a core skill for critical thinking and decision-making. Statistics doesn't just organize numbers; it uncovers patterns, quantifies uncertainty, and transforms raw information into knowledge we can act on.

This module takes learners on a journey from the basics to the essentials of statistical reasoning. We start with data types and collection methods, then move to how data can be organized and presented through clear tables, visuals, and descriptive summaries. We dive into measures of central tendency and dispersion to understand what data is really telling us, before laying the groundwork of probability and distributions as the language of uncertainty.

From there, learners will explore statistical inference, confidence intervals and hypothesis testing to make evidence-based generalizations from samples to populations. By the end, participants won't just know statistical methods; they'll be able to apply them confidently, communicate insights clearly, and make better decisions in real-world contexts.

# Preface

## About the Writer



[Bakti Siregar, M.Sc., CDSS](#) 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

In an era dominated by data, mastering statistics is crucial for making evidence-based decisions and revealing meaningful patterns within complex datasets. This module

introduces learners to the fundamental principles and methods of statistics, equipping them with the skills to explore, summarize, and interpret data effectively. This Book covers:

- Introduction to statistics and its role in decision-making
- Data types and collection methods for accurate and reliable analysis
- Data presentation using clear tables, charts, and visual summaries
- Measures of central tendency and dispersion to describe datasets
- Probability concepts and probability distributions to quantify uncertainty
- Confidence intervals and statistical inference for drawing robust conclusions
- Nonparametric methods for analyzing data without strict distribution assumptions

By completing this module, learners will gain the analytical capabilities to manage real-world data, extract actionable insights, and communicate findings with clarity and rigor, establishing a strong foundation for advanced study or professional practice in data science, research, and industry.

---

## Feedback & Suggestions

Your feedback is essential for improving the clarity, relevance, and usefulness of this module. Readers are invited to share their thoughts on the content, structure, and practical applications, as well as suggestions for new topics, examples, or tools.

This input helps make the E-book a more practical and comprehensive resource for **Introduction to Statistics**, bridging academic learning and real-world application. Thank you for contributing to the evolution of this material!

For feedback and suggestions, feel free to contact:

- dsciencelabs@outlook.com

- siregarbakti@gmail.com

- siregarbakti@itsb.ac.id

# About R and RStudio

R and RStudio are essential tools for data analysis, statistical computing, and visualization. R provides a powerful, open-source environment for performing complex analyses, while RStudio offers a user-friendly interface, supporting multiple languages and features for coding, documentation, and reproducible research. Mastery of R and RStudio enables users to explore data efficiently, implement statistical methods, and communicate insights effectively in scientific, engineering, business, and research contexts [1], [2].

The Figure 1 presents a visual overview of this introductory material, highlighting the main topics—R, RStudio, Installation, Usage, and Popularity—and their subtopics. It serves as a roadmap for readers, showing how foundational knowledge of R and RStudio connects to practical applications, package management, data analysis workflows, and understanding the broader statistical and computational ecosystem [3], [4].
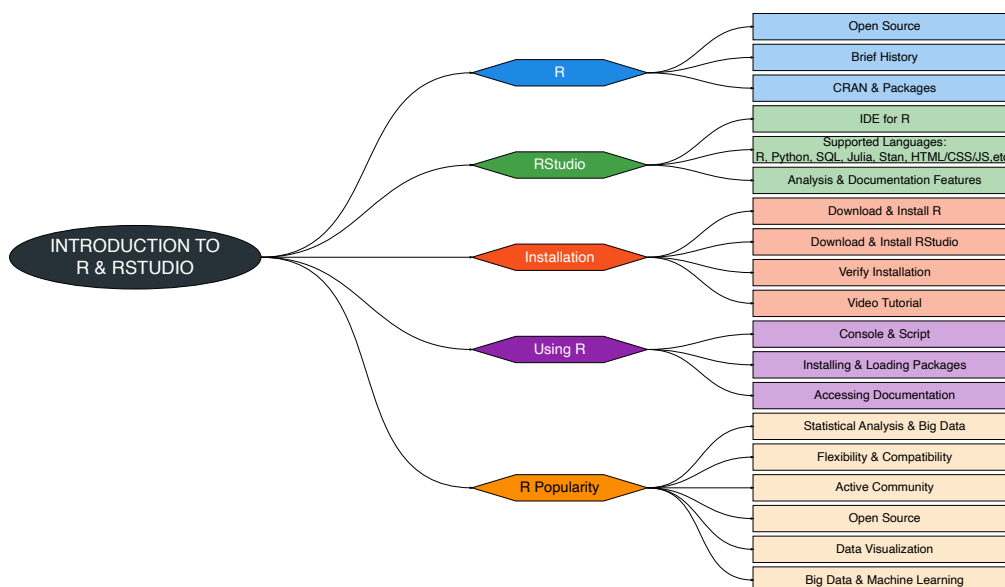


Figure 1: Mind Map of Introduction to R & RStudio

The mind map above (Figure Figure 1) provides a structured overview of the core topics in this chapter: **R**, **RStudio**, **Installation**, **Usage**, and **Popularity**. Each branch and sub-branch highlights essential concepts and practical steps, showing how

they interconnect to form a complete understanding of statistical computing and data analysis workflows. By following this visual roadmap, readers can see how mastering the fundamentals of R and RStudio—from installing software and running basic scripts to exploring packages and advanced features—lays the groundwork for effective data analysis, reproducible research, and real-world problem solving. This chapter will guide you step by step through each component, linking theory to hands-on applications and best practices.

# Introduction to R & RStudio

R and RStudio are open-source applications widely used in big data and data science. The combination of both allows users to perform complex data analysis and visualization efficiently and easily.

These applications are examples of open-source software, meaning they can be freely used, modified, and distributed. More information about open-source software can be found here: What is Open Source Software?

## Brief History of R

The R programming language (Figure 2) was developed in the early 1990s by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. The goal was to create a better data analysis tool than other statistical languages such as S. R was released in 1995 and quickly gained attention from the statistical community.

As an open-source language, R grew rapidly with global contributions. CRAN (Comprehensive R Archive Network), founded in 1997, provides thousands of community packages extending R's functionality. R's popularity increased in the early 2000s, expanding into industry and academia.

## About RStudio

Launched on February 21, 2011, Figure 3 was founded by J.J. Allaire, also known for his role in early web technologies such as ColdFusion. RStudio has become one of the most popular IDEs for R, offering many features to facilitate data analysis, coding, and dynamic documentation using R Markdown.

RStudio supports multiple programming languages:

- **R**: Primary language for data analysis.
- **Python**: Via `reticulate` for data analysis.
- **SQL**: With `DBI` package for database queries.
- **Stan**: Via `rstan` for Bayesian modeling.
- **Julia**: With `JuliaCall` for high-performance computing.
- **Shell (Bash)**: For system commands in the terminal.
- **HTML/CSS/JavaScript**: In R Markdown for web documents.

Figure 2: Logo R



Figure 3: Logo RStudio

# Installing R and RStudio

## Step 1: Download and Install R

**Download R:**

- Visit CRAN R
- Select "Download R for Windows" (or your OS)
- Click "base" to get the latest version
- Download the installer according to your system (32-bit or 64-bit)

**Install R:**

- Run the downloaded installer
- Follow on-screen instructions
- Choose installation directory if needed
- Click "Finish" when done

**Note:** Ensure **R** is correctly installed before proceeding to RStudio.

## Step 2: Download and Install RStudio

**Download RStudio:**

- Visit RStudio
- Select "RStudio Desktop"
- Download the free version ("RStudio Desktop Open Source License") or paid version as needed

**Install RStudio:**

- Run the installer
- Follow on-screen instructions
- Choose installation directory if needed
- Click "Finish" when done

## Step 3: Verify Installation

**For R:**

- Open R from Start menu or desktop
- Type `version` in console and press Enter
- Ensure the version displayed is up to date

**For RStudio:**

- Open RStudio
- Check that it connects to the installed R
- Run basic commands like `2 + 2` to ensure functionality

# Installation Video

\newline \href{https://youtu.be/Lv0xcdeXaGU}{Click here to watch the video}

# Popularity of R

R is widely recognized among data scientists and researchers. Key reasons for its popularity include:

## Statistical Analysis and Big Data

R is efficient for statistical and big data analysis (Figure 4) thanks to many supporting packages and libraries.



Figure 4: Dashboard Example

## Flexibility and Compatibility

R is flexible and compatible (Figure 5) with multiple platforms, making integration with other software easy.

Figure 5: Flexibility and Compatibility

## Active Community

R has a large, active user community providing resources for learning and sharing knowledge.

- **R Project**: Official site
- **Mailing Lists**: Subscribe for updates about R releases here
- **Twitter #rstats**: Active users share insights on Twitter link
- **Tidy Tuesday**: Weekly online project for data visualization with open-source datasets link
- **R-Ladies**: Global group promoting gender equality in R community link
- **R-Podcast**: Podcast with R tips and updates link
- **R-Bloggers**: Blog site for sharing R code, analysis, and visualization link

## Open Source

As open-source software, R can be freely used and developed, making it ideal for researchers with limited budgets (See Figure 6).

## Data Visualization

R excels in data visualization (Figure 7), presenting complex data clearly and attractively.

Figure 6: Open Source
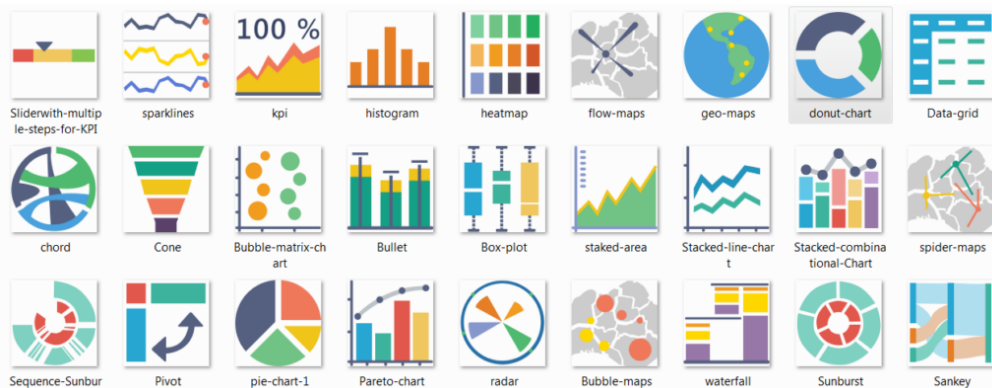


Figure 7: Data Visualization

**Suitable for Big Data & ML**

As the world of data grows larger and more complex, R keeps pace by offering tools designed for big data and Machine Learning (ML). This Figure 8 highlights R's strength in combining its statistical roots with modern capabilities, enabling analysts, researchers, and businesses to explore data, build models, and generate insights with confidence.



Figure 8: Big Data & Machine Learning

# How to Use R/Studio

To start using R effectively, follow these steps:

- **R:** Open the R application from Start menu or desktop to access the console.
- **RStudio:** Open RStudio for a graphical interface that simplifies coding and analysis.

## Writing and Running Code

- Klik **Console Tab:** Enter commands directly in "Console", Example:

```
print("Hello, World!")
```

- **Script Tab:** Save and run multiple commands, Example:

```
# Simple R script
x <- 10
y <- 5
result <- x + y
print(result)
```

## Installing and Loading Packages

- **Install Packages:**

```
install.packages("ggplot2")
```

- **Load Packages:**

```
library(ggplot2)
```

## Accessing Documentation

- **Function Help:**

```
help(plot)
?plot
```

- **Vignettes:**

```
vignette("ggplot2")
```

# About the Book

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make informed decisions. It provides essential tools for understanding variability, modeling uncertainty, and drawing conclusions from real-world phenomena across science, engineering, business, and social studies. Mastery of statistics enables us to extract insights, test hypotheses, and predict outcomes effectively [5], [6].

## Overview of the Course

The Figure 9 presents a visual overview of the course, highlighting the structure of key topics and their interconnections. It offers readers a clear guide to navigate the material and understand how concepts link to practical applications and decision-making processes [7].

This book introduces the fundamental building blocks of statistics, from understanding data structures and basic visualizations to exploring probability, distributions, confidence intervals, and nonparametric methods. Each topic is linked to real-world examples, allowing readers to see how statistical techniques support analysis, interpretation, and problem-solving across diverse domains.

## Brief Descriptions

This mind map (Figure 9) illustrates the overall structure of a Basic Statistics course, covering topics from introductory concepts to more advanced methods (see Table 1).
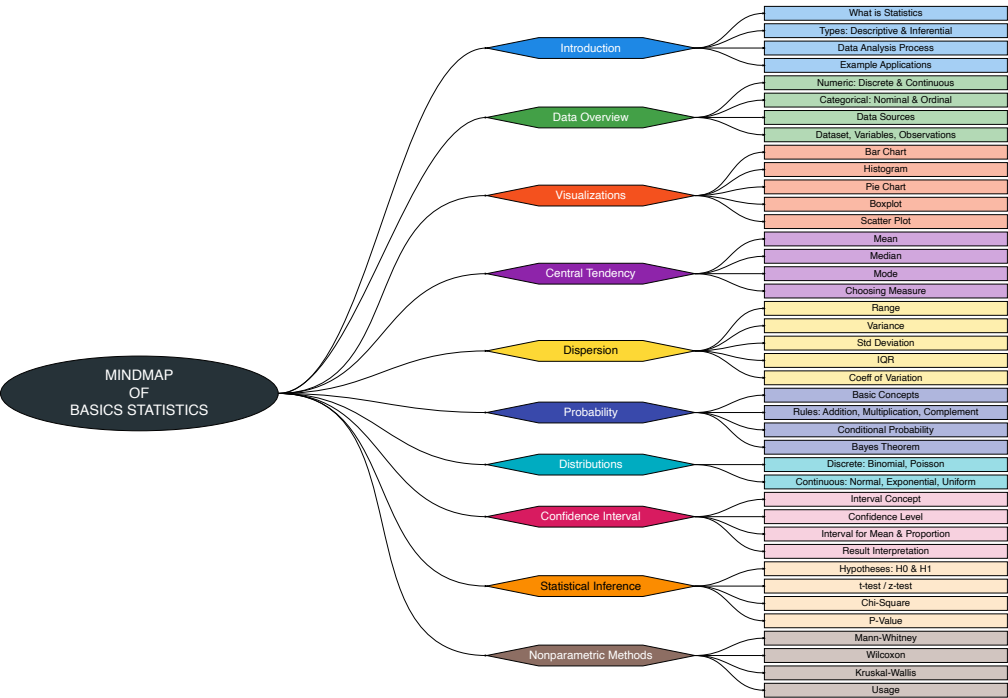
Figure 9: Mind Map of Statistics Course

Table 1: Key Concepts in Statistics

| KeyConcept | Description | ExampleApplication |
|---|---|---|
| Introduction | What statistics is, types (descriptive & inferential), and the data analysis process | Business decision-making using data insights |
| Data Overview | Types of data (numerical, categorical), data sources, datasets, variables, and observations | Collecting employee health records for analysis |
| Visualizations | Visualization techniques: bar chart, histogram, pie chart, boxplot, scatter plot | Visualizing sales data with bar chart or boxplot |
| Central Tendency | Measures of location: mean, median, mode | Comparing average income across groups |
| Dispersion | Measures of variability: range, variance, standard deviation, IQR, coefficient of variation | Analyzing spread of exam scores in a class |
| Probability | Basic concepts, rules (addition, multiplication), conditional probability, Bayes' theorem | Estimating probability of machine failure |
| Distributions | Discrete (binomial, Poisson) and continuous (normal, exponential, uniform) distributions | Modeling customer arrivals (Poisson) or product lifespan (exponential) |
| Confidence Interval | Intervals, confidence levels, estimation for mean & proportion, interpretation of results | Calculating CI for average mining output |
| Statistical Inference | Hypothesis testing (H0 & H1), t-test, z-test, chi-square, p-values | Testing if two mining methods yield different results |
| Nonparametric Methods | Mann-Whitney, Wilcoxon, Kruskal-Wallis tests, and when to use them | Analyzing survey responses when assumptions of parametric tests are not met |

# Chapter 1

# Intro to Statistics

Statistics appears in almost every aspect of daily life. When reading news reports about surveys, public health updates, or economic analysis, we are already looking at applications of statistics. It helps us transform raw data into meaningful information that supports better understanding and decision-making. This chapter introduces the meaning of statistics, its main types, the process of data analysis, and practical applications across different fields.

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make informed decisions. It provides essential tools for understanding variability, modeling uncertainty, and drawing conclusions from real-world phenomena across science, engineering, business, and social studies. Mastery of statistics enables us to extract insights, test hypotheses, and predict outcomes effectively [5], [6].

The Figure 1.1 presents a visual overview of the course, highlighting the structure of key topics and their interconnections. It offers readers a clear guide to navigate the material and understand how concepts link to practical applications and decision-making processes [7].

Statistics is a fundamental discipline in data science, serving as a foundation for understanding, analyzing, and interpreting information. By applying the 5W+1H framework (What, Why, When, Where, Who, How), we can systematically explore the essence of statistics: its definition, purpose, history, areas of application, contributors, and methodology.

Table Table 1.1 provides an overview of these guiding questions, linking each with practical examples and interpretations that reflect both everyday understanding and scientific perspectives.

## 1.1 Definition of Statistics

### 1.1.1 The Meaning of Statistics

**Everyday explanation:** Statistics is a way of making data easier to understand. Imagine a teacher who wants to know how well the class performed on an exam. Instead

Table 1.1: 5W+1H Questions for Statistics

| | Description | Example_Stat | Example_Output |
|---|---|---|---|
| **What?** | | | |
| What? | What is statistics? | Science of collecting, organizing, analyzing, and interpreting data | Tool to make sense of uncerta |
| What? | What are the main branches of statistics? | Descriptive and Inferential statistics | Descriptive: summarize data; Inferential: draw conclusions |
| What? | What is the role of data in statistics? | Data as the raw material for statistical inference | Without data, no statistical inference is possible |
| **Why?** | | | |
| Why? | Why is statistics important for decision-making? | Helps reduce uncertainty and guide policies | Example: public health decis during a pandemic |
| Why? | Why do we use statistics in research and business? | To validate research findings, optimize business strategies | Example: forecasting sales, testing medical treatments |
| **When?** | | | |
| When? | When did statistics begin to be formalized? | 18th–19th century (Gauss, Laplace, Fisher, Pearson) | Roots in census-taking, formalized with probability theory |
| When? | When is statistical analysis applied in practice? | Market research, medical studies, social surveys | Example: analyzing customer satisfaction survey |
| **Where?** | | | |
| Where? | Where is statistics applied in real-world problems? | Business, economics, health, engineering, social sciences | Example: clinical trials, risk assessment, AI systems |
| Where? | Where can statistical thinking be observed in daily life? | Everyday: opinion polls, product reviews, budgeting | Example: choosing insurance plans, election predictions |
| **Who?** | | | |
| Who? | Who developed the foundations of modern statistics? | Key figures: Ronald Fisher, Karl Pearson, Florence Nightingale | Pioneers advanced probability statistical theory |
| Who? | Who uses statistics in professional fields? | Researchers, policy makers, engineers, doctors, data scientists | Used across all scientific and professional domains |
| **How?** | | | |
| How? | How is data collected in statistics? | Surveys, experiments, sensors, digital footprints | Quantitative and qualitative sources |
| How? | How is data analyzed and modeled? | Using EDA, hypothesis testing, regression, machine learning | Models patterns, tests hypotheses, builds prediction |
| How? | How are results interpreted and communicated? | Through reports, dashboards, visualizations, publications | Translate numbers into meaningful insights |
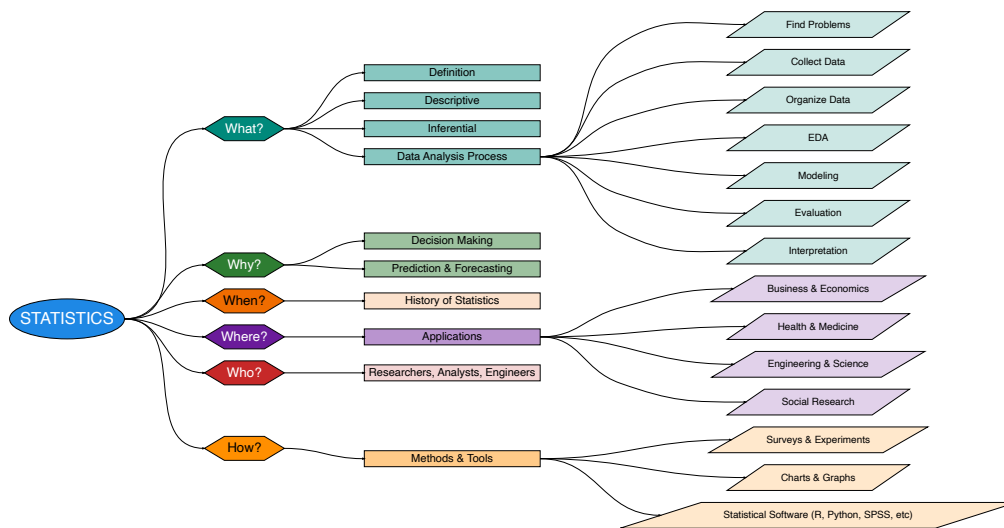
Figure 1.1: Detailed 5W+1H for Statistics

of looking at every student's score one by one, the teacher can simply calculate the average score to get an overall picture.

**Scientific explanation:** Statistics is a branch of mathematics concerned with the methods of **collecting, organizing, analyzing, interpreting, and presenting data**. Its main purpose is to turn raw observations into reliable information for reasoning and decision-making.

> **Example:**
> Raw scores: [65, 70, 75, 80, 90]
> Descriptive result: mean = 76, median = 75
> Conclusion: The class average is fairly good.

## 1.1.2 Statistics in Decision-Making

Statistics is especially valuable when decisions must be made under uncertainty. A shop owner might record daily sales to decide which day is best for restocking. A doctor may evaluate the effectiveness of a new treatment by analyzing patient data.

**In academic terms,** statistics supports:

- summarizing large datasets,

- identifying relationships among variables,

- predicting future outcomes,

- and enabling **evidence-based decisions**.

## 1.2   Types of Statistics

### 1.2.1   Descriptive Statistics

Descriptive statistics focuses on **summarizing and presenting data** in a meaningful way.  It includes measures of central tendency (mean, median, mode), measures of variability (variance, standard deviation, range), and visualization tools like tables, histograms, and boxplots.

> **Example:** From 100 students, the average exam score is 72, the highest is 95, and the lowest is 40.  A histogram shows how scores are distributed across the group.

### 1.2.2   Inferential Statistics

Inferential statistics goes beyond description. It aims to make **generalizations about a population** based on data from a smaller sample.

> **Example:** A sample of 100 students has an average score of 72.  Using inferential techniques, we estimate that the average score of the entire university (10,000 students) lies between 71 and 73 with 95% confidence.

Common methods include hypothesis testing, confidence intervals, regression analysis, and ANOVA.

## 1.3   Data Analysis Process

Before we go further, let's take a moment to watch a short video about statistics. This vidoe bellow will help you see how statistics is used in everyday life and why it is so important in many fields. By watching it, you will get a clearer picture of how numbers and data can guide decisions, solve problems, and make our world easier to understand.

\newline \href{https://youtu.be/Lv0xcdeXaGU}{Click here to watch the video}

Analyzing data involves several stages, each building upon the previous one.  This process ensures that the final conclusion is accurate and meaningful.

1. **Defining the Problem**
   The process begins with a clear question. For example: *Does online advertising increase sales?*

2. **Collecting Data**
   Data can be obtained through surveys, experiments, observations, or secondary sources such as databases and official reports.

3. **Organizing Data**
   Raw data is often messy. This step includes cleaning errors, removing duplicates, handling missing values, and structuring the data in tables.

4. **Exploratory Data Analysis (EDA)**
   Before modeling, data is explored to identify distributions, trends, or outliers. Visual tools like scatter plots or boxplots are particularly useful here.

5. **Modeling**
   Statistical or machine learning models are applied to draw deeper insights. Linear regression predicts outcomes, classification assigns groups, and time series analysis forecasts future values.

6. **Evaluating the Model**
   Models are tested for accuracy. Regression models use $R^2$ or RMSE, while classification models rely on accuracy, precision, recall, and F1-score.

7. **Interpreting Results**
   Numbers are translated into real-world meaning. For example: *Every additional $1,000 spent on advertising is associated with an increase of 50 sales units.*

## 1.4   Applied of Statistics

Statistics is not only a theoretical field but also a discipline with wide-ranging applications across real-world domains. Its methods enable decision-making, provide evidence-based insights, and support the development of new knowledge in many sectors. Whether in the corporate world, medical research, engineering innovations, or social sciences, statistics acts as a bridge between raw data and meaningful conclusions.

Table Table 1.2 highlights several key areas where statistics is applied, explaining the role it plays and offering concrete examples that demonstrate its importance in practice.

## References

Table 1.2: Applications of Statistics in Different Fields

|                          | Explanation                                                                                              | Illustrative Example                                            |
|--------------------------|----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|
| Business and Economics   | Companies use statistics to analyze sales trends, forecast demand, set prices, and manage investment risks. | Example: Predicting next quarter sales or assessing portfolio risk. |
| Health and Medicine      | Statistical methods guide clinical trials, monitor disease spread, and evaluate the effectiveness of treatments. | Example: Testing a new vaccine for safety and efficacy.         |
| Engineering and Science  | Engineers and scientists apply statistics to quality control, material testing, experimental design, and environmental modeling. | Example: Evaluating durability of construction materials.       |
| Social Research          | Governments and researchers rely on statistics for population surveys, educational assessments, and policy evaluation. | Example: Using census data to design social welfare programs.   |

# Chapter 2

# Data Overview

## 2.1 Types of Data

## 2.2 Numeric (Quantitative)

### 2.2.1 Discrete

### 2.2.2 Continuous

## 2.3 Categorical (Qualitative)

### 2.3.1 Nominal

### 2.3.2 Ordinal

## 2.4 Data Sources

## 2.5 Data Structure

### 2.5.1 Dataset

### 2.5.2 Variables

### 2.5.3 Observations

## References

# Chapter 3

# Basic Data Visualizations

# Chapter 4

# Central Tendency

# Chapter 5

# Statistical Dispersion

# Chapter 6

# Essentials of Probability

# Chapter 7

# Probability Distributions

## 7.1  Discrete

### 7.1.1  Binomial

### 7.1.2  Poisson

## 7.2  Continuous

### 7.2.1  Normal

### 7.2.2  Exponential

### 7.2.3  Uniform

# Chapter 8

# Confidence Interval

**8.1   Interval Concept**

**8.2   Confidence Level**

**8.3   Interval for Mean & Proportion**

**8.4   Result Interpretation**

# Chapter 9

# Statistical Inference

## 9.1   Statistical Hypotheses

### 9.1.1   Null Hypothesis (H )

### 9.1.2   Alternative Hypothesis (H )

## 9.2   T-test, Z-test

## 9.3   Chi-Square Test

## 9.4   P-Value and Decision Making

# Chapter 10

# Nonparametric Methods

## 10.1  Mann-Whitney

## 10.2  Wilcoxon

## 10.3  Kruskal-Wallis

## Usage

[1]   Ihaka, R. and Gentleman, R., R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, 299–314, 1996

[2]   Allaire, J. J., RStudio: Integrated development environment for r, https://rstudio.com/, 2011

[3]   Comprehensive r archive network (CRAN), https://cran.r-project.org/, 1997

[4]   RStudio documentation and resources, https://rstudio.com/, 2021

[5]   Moore, D. S., McCabe, G. P., and Craig, B. A., The practice of statistics, W.H. Freeman; Company, New York, 2020

[6]   Wackerly, D. D., Mendenhall, R., and Scheaffer, R., Mathematical statistics with applications, Cengage Learning, Boston, 2014

[7]   Freedman, D., Pisani, R., and Purves, R., Statistics, W.W. Norton & Company, New York, 2007