ITSB | Eco Industry Oriented University

sinarmas

# INTRODUCTION TO STATISTICS

## A Data Science Perspective with R


**Written by:**

Bakti Siregar, M.Sc., CDS.

D'SCIENCELABS
Smart Idea

Kampus Merdeka
INDONESIA JAYA

First Edition

# 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., CDS is a Lecturer in the Data Science Program at ITSB. He obtained his Master's degree in Applied Mathematics from the National Sun Yat-sen University, Taiwan. Alongside his academic role, Bakti also serves as a Freelance Data Scientist, collaborating with leading companies such as JNE, Samora Group, Pertamina, and PT. Green City Traffic.

His professional and research interests include Big Data Analytics, Machine Learning, Optimization, and Time Series Analysis, with a particular focus on finance and investment applications. His core expertise lies in statistical programming using R and Python, complemented by strong experience in database management systems such as MySQL and NoSQL. In addition, he is proficient in applying Big Data technologies, including Spark and Hadoop, for large-scale data processing and analysis.

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].

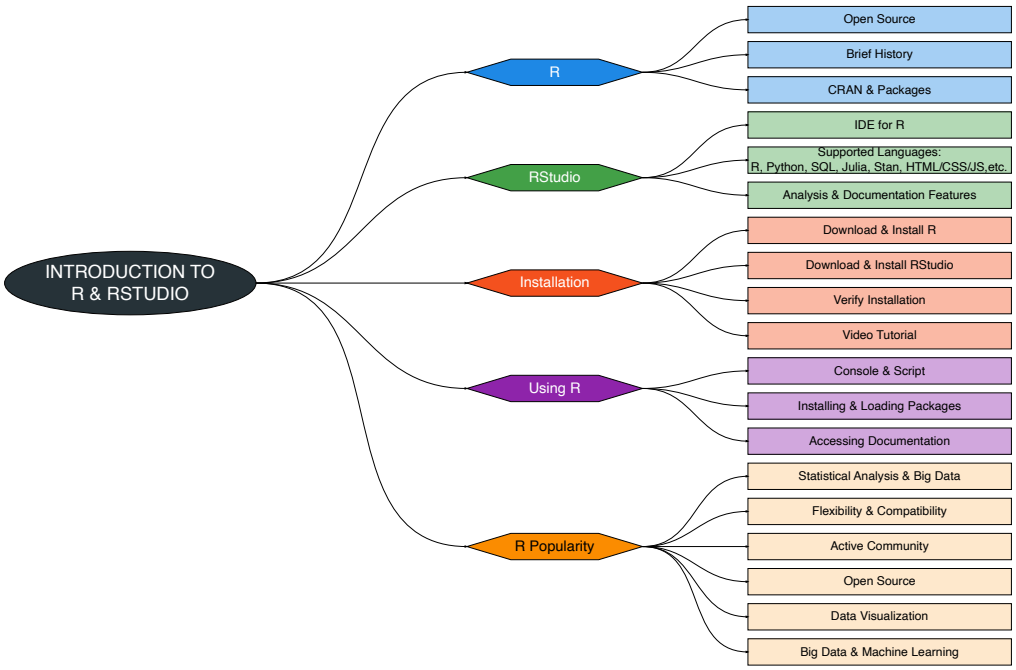



Figure 1: Mind Map of Introduction to R & RStudio

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.

Figure 2: Logo R

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.

Figure 3: Logo RStudio

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.

# 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.

## Flexibility and Compatibility

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

## 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 4: Dashboard Example


Figure 5: Flexibility and Compatibility

### 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.

## 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:


Figure 6: Open Source


Figure 7: Data Visualization


Figure 8: Big Data & Machine Learning

```
# 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")
```

# References

# 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].

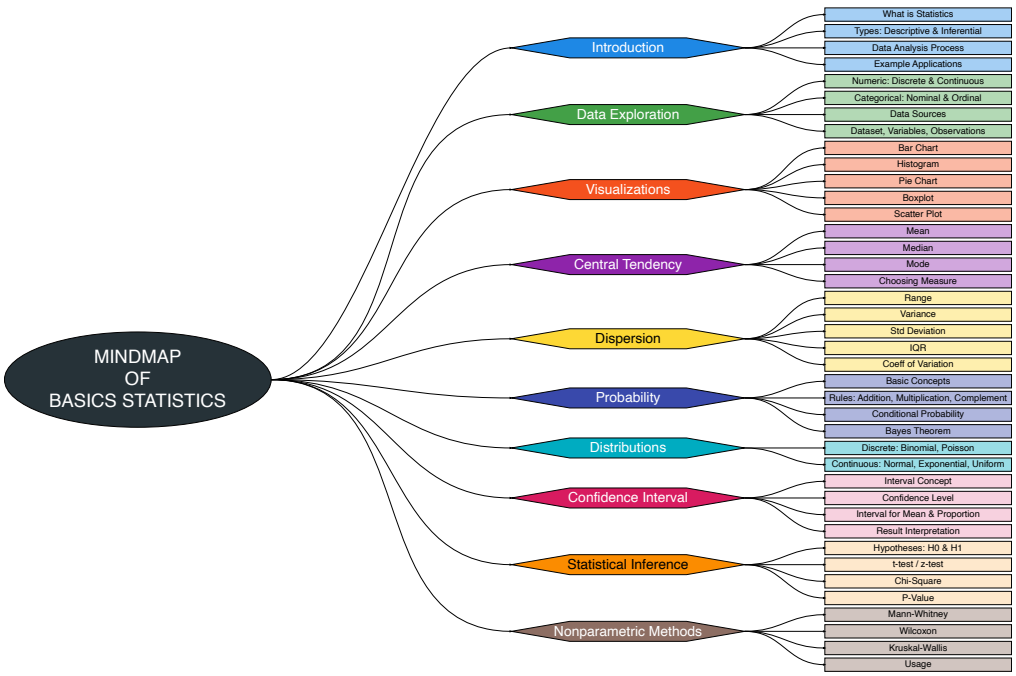



Figure 9: Mind Map of Statistics Course

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 Exploration | 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 |

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).

## References

# Chapter 1

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

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 uncertainty |
| 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 inferen is possible |
| **Why?** | | | |
| Why? | Why is statistics important for decision-making? | Helps reduce uncertainty and guide policies | Example: public health decisions during a pandemic |
| Why? | Why do we use statistics in research and business? | To validate research findings, optimize business strategies | Example: forecasting sales, testin medical treatments |
| **When?** | | | |
| When? | When did statistics begin to be formalized? | 18th–19th century (Gauss, Laplace, Fisher, Pearson) | Roots in census-taking, formalize 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 plan 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 data sources |
| How? | How is data analyzed and modeled? | Using EDA, hypothesis testing, regression, machine learning | Models patterns, tests hypotheses, builds predictions |
| How? | How are results interpreted and communicated? | Through reports, dashboards, visualizations, publications | Translate numbers into meaningfu insights |

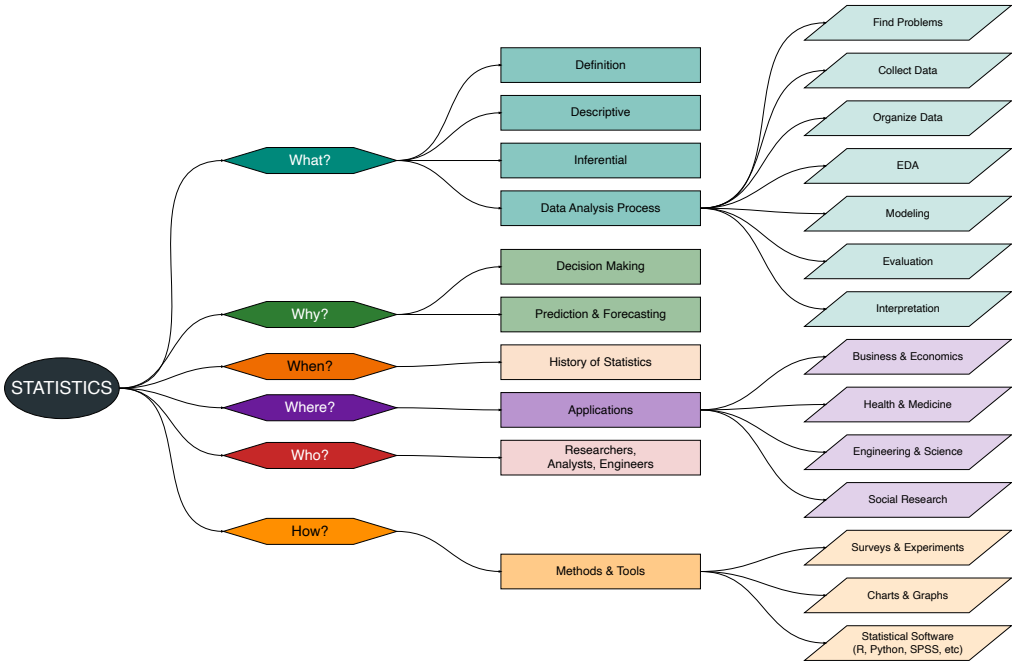



Figure 1.1: Detailed 5W+1H for Statistics

**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.

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. |

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

# Chapter 2

# Data Exploration

After understanding the important role of statistics in turning raw data into meaningful insights as mentioned in chapter Intro to Statistics, the next step is to explore **the nature of data** and how it can be classified. Data forms the foundation of any analysis, and without a clear understanding of its types and structure, organizing, interpreting, and making accurate decisions can be challenging [8].

This section provides a **Data Exploration** Figure 2.1, covering the classification of data into **numeric (quantitative)** and **categorical (qualitative)** types, including subtypes such as *discrete, continuous, nominal,* and *ordinal* [9]. It also discusses **data sources** and the basic structure of a dataset, including *variables* and *observations* [10]. By mastering these concepts, readers will gain a solid foundation for subsequent analytical steps and will be better equipped to recognize and handle different forms of data in context [11]–[15].

## 2.1 Types of Data

In statistics, understanding the types of data is a crucial starting point. Data can be broadly divided into two main groups: numerical and categorical. Numerical data represent numbers that can be either discrete (countable, such as the number of students) or continuous (measurable, such as height or temperature) [16]. Categorical data, on the other hand, represent labels or groups. They can be nominal (without order, such as gender or colors) or ordinal (with order, such as satisfaction levels: low, medium, high) [17].

Knowing the correct type of data is essential because it guides us in choosing the right statistical methods, the most suitable visualizations, and ensures that our interpretations are accurate [18]. The following video will help you clearly understand these concepts through simple explanations and real-world examples.

**Watch here:** Types of Data — Categorical vs Numerical

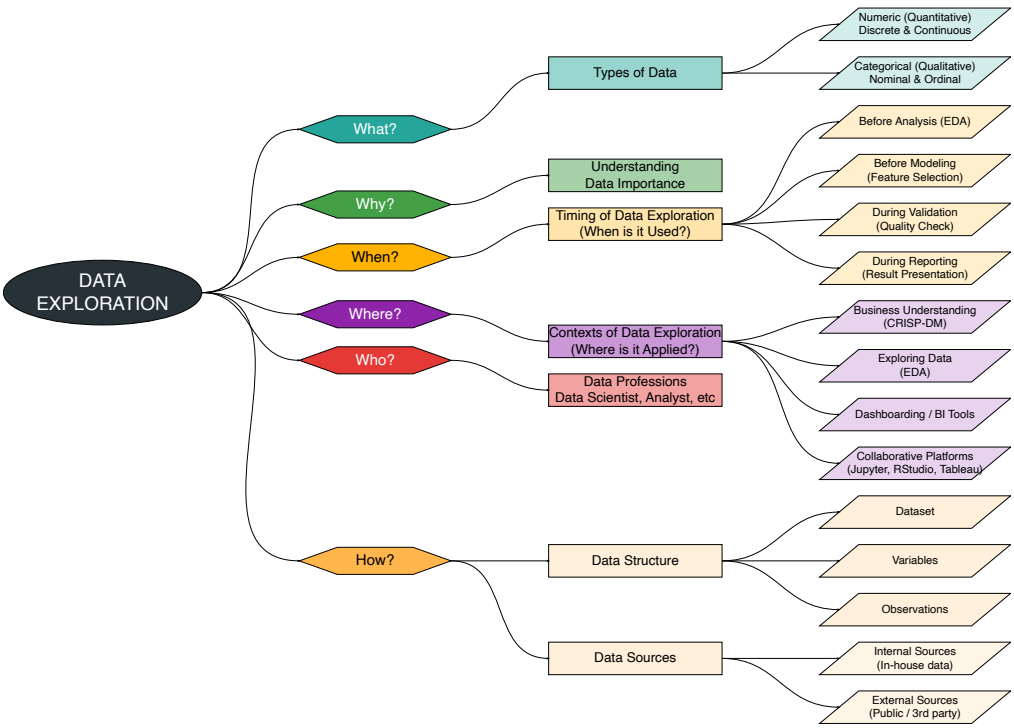



Figure 2.1: Data Exploration 5W+1H

## 2.2 Numeric (Quantitativ)

Numeric or quantitative data are data expressed in numbers that represent counts or measurements[9]. They provide information about **how much** or **how many** of something, allowing for mathematical operations such as addition, subtraction, averaging, and statistical analysis [10].

Quantitative data are divided into two main types:

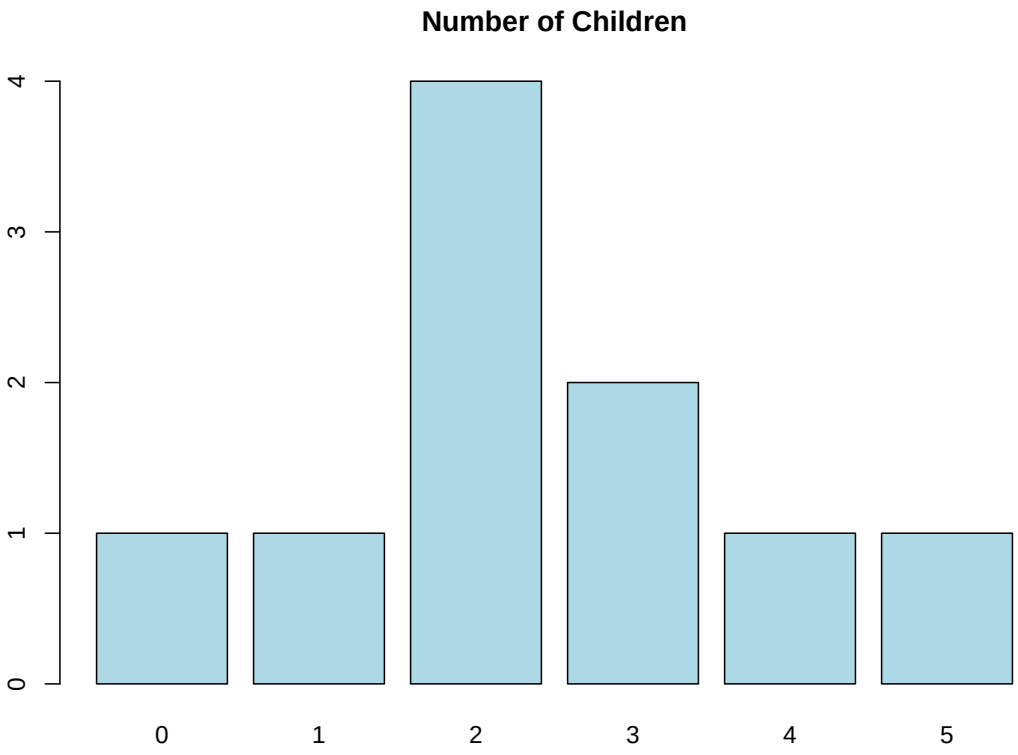
- **Discrete data**: consist of countable whole numbers (e.g., number of students, number of cars) [8].
- **Continuous data**: consist of measurable values that can take on decimals (e.g., height, weight, temperature) [9].

### 2.2.1 Discrete

Discrete data are numerical values that can be counted and usually take whole numbers [8], [9]. These data cannot contain fractions or decimals, since each value represents a complete count. Examples include the number of children in a family, the number of cars owned, or the number of accidents in a month [8].

```r
children <- c(2, 3, 1, 4, 2, 3, 2, 5, 0, 2) # Discrete Data Example
children                                     # Print result (way 1)
print(children)                              # Print result (way 2)
table(children)                              # frequency distribution
mean(children)                               # average
```

```r
# Basic Visual
barplot(table(children),
        main="Number of Children",
        col="lightblue")
```

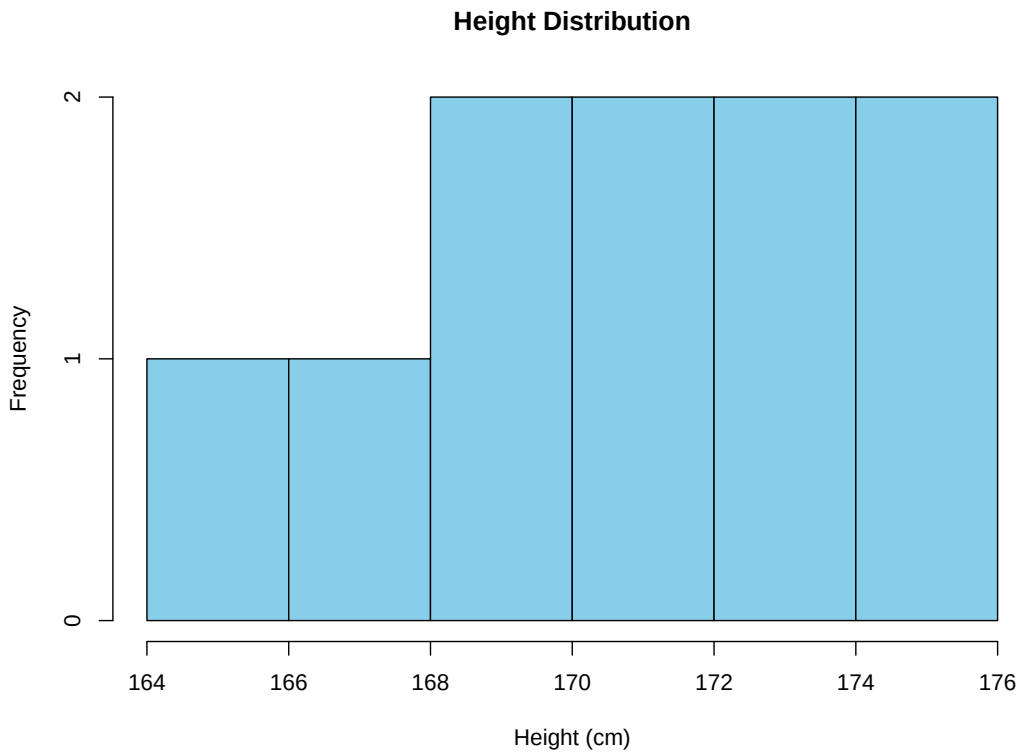



### 2.2.2 Continuous

Continuous data are numerical values obtained through measurement and can include fractions or decimals [9], [10]. These values are not limited and can take on any value within a given range. Examples include height, weight, temperature, and rainfall [9].

```
# Continuous Data Example
height <- c(165.2, 170.5, 172.3, 168.8, 174.1,
            169.4, 171.7, 173.6, 175.2, 166.8)
summary(height)
```

```
hist(height,
     col="skyblue",
     main="Height Distribution",
     xlab="Height (cm)")
```




## 2.3 Categorical (Qualitative)

Categorical or qualitative data are data expressed in labels, names, or categories rather than numbers [9]. They describe **qualities, attributes, or classifications** that cannot be meaningfully measured with arithmetic operations like addition or subtraction.

Categorical data are divided into two main types:

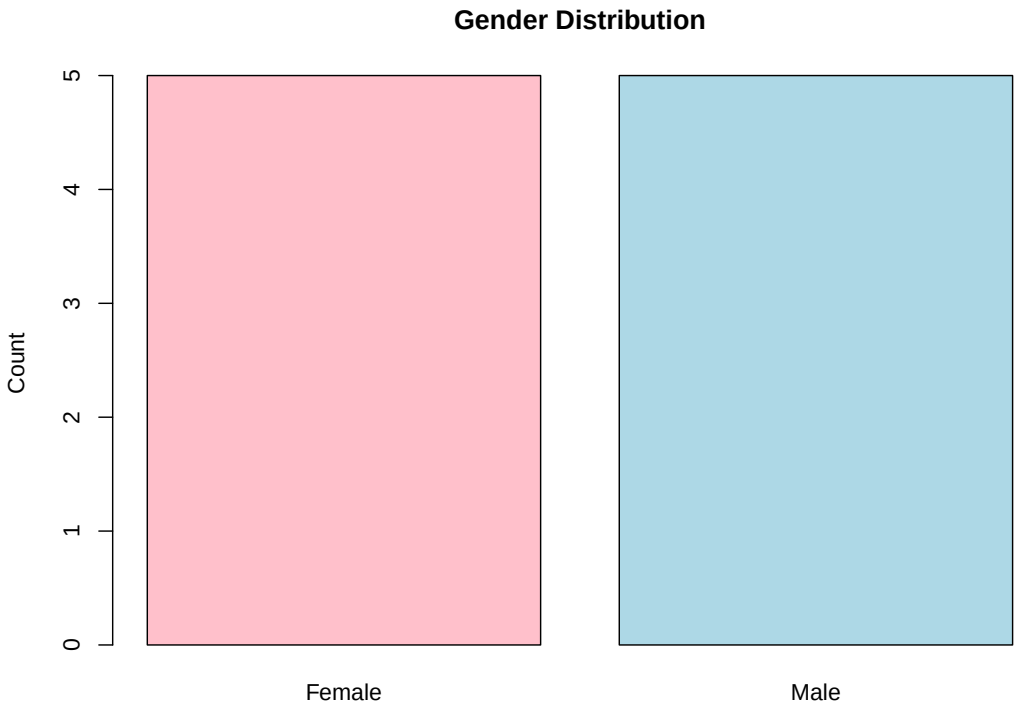
- **Nominal data**: categories without any natural order or ranking (e.g., gender, blood type, car brand) [8].
- **Ordinal data**: categories with a meaningful order or ranking, but without fixed differences between ranks (e.g., education level, satisfaction rating, socioeconomic status) [10].

### 2.3.1 Nominal

Nominal data are categorical values that act only as labels or identifiers, with no inherent order or ranking [8], [9]. They are used to classify objects into different groups, but there is no meaning of greater or lesser among the categories. Examples include gender, blood type, and product brands.

```r
# Nominal Data Example
gender <- c("Male", "Female", "Male", "Male", "Female", "Female",
            "Male", "Female", "Male", "Female")
table(gender)
```

```r
barplot(table(gender),
        col=c("pink","lightblue"),
        main="Gender Distribution",
        ylab="Count")
```

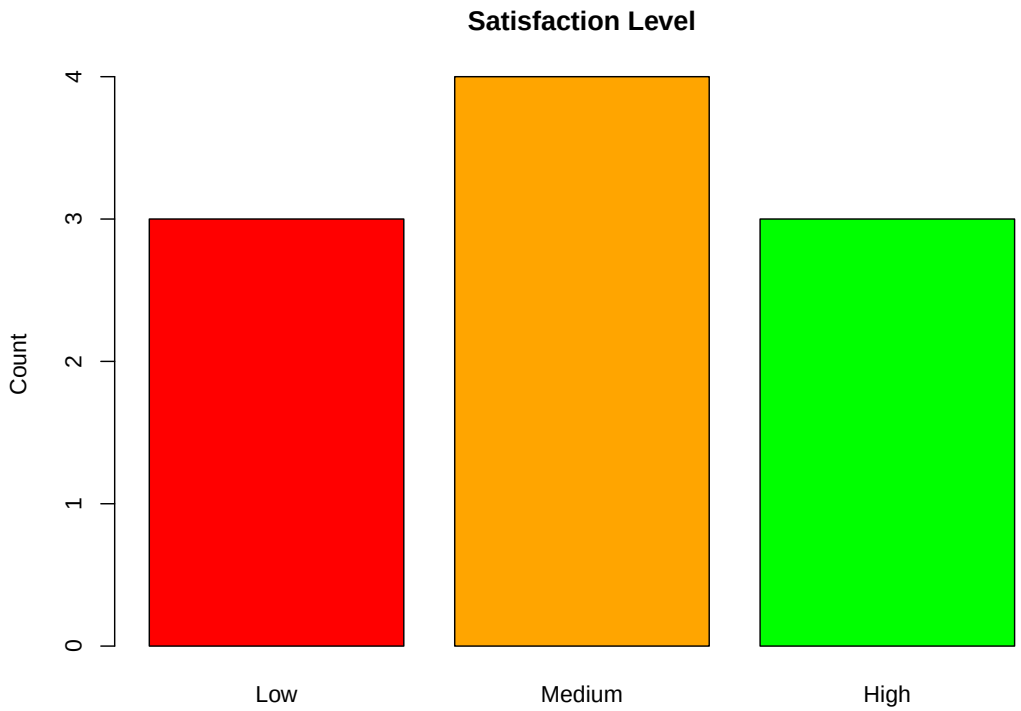



### 2.3.2 Ordinal

Ordinal data are categorical values that have a clear order or ranking, but the distance between categories is not precisely measurable [9]. These data show levels or rankings but do not indicate the magnitude of differences between them. Examples include satisfaction levels (low, medium, high), education levels, or competition rankings [9].

```r
# Ordinal Data Example
satisfaction <- factor(c("Low","Medium","High","Medium","High","Low",
                         "Medium","High","Medium","Low"),
                       levels = c("Low","Medium","High"), ordered = TRUE)

table(satisfaction)
```

```
barplot(table(satisfaction),
        col=c("red","orange","green"),
        main="Satisfaction Level",
        ylab="Count")
```




## 2.4 Data Sources

**Data Sources** are the origins of data used for analysis. Knowing the source is important because it affects **data quality, validity, and relevance** [19].

**Watch here:** Handling your Data Sources

**Types of Data Sources:**

1. **Internal Sources** – Data coming from within the organization, e.g., sales transactions, inventory records, financial reports, or employee data [20].
2. **External Sources** – Data obtained from outside the organization, e.g., government statistics, industry reports, public datasets, social media, or third-party providers [20].
3. **Structured vs Unstructured Data**
    - **Structured Data:** Organized in tables or databases, easy to analyze [19].
    - **Unstructured Data:** Text, images, videos, or log files that require preprocessing [19].

Consider the the following Video to know more about Structured and Unstructured Data!.

**Watch here:** Structured vs Unstructured Data

## 2.5 Data Structure

**Data Structure** refers to the way data is organized to make analysis easier and more efficient. A well-structured dataset helps with **cleaning, processing, analyzing, and visualizing data** [19]; [20]. The main components of data structure are:

- **Dataset:** A collection of data arranged in a structured format, usually as a table.
- **Columns:** Each column represents a **variable** or attribute describing the observations.
- **Rows:** Each row represents a single **observation** or case.

### 2.5.1 Dataset (Data Frame)

**Example:** An online store wants to analyze its sales performance over the first week of October 2025. They collect the following information for each transaction:

| Column | Type | Description |
|---|---|---|
| Date | Date | The date of the transaction |
| Qty | Discrete | The quantity sold (countable numbers) |
| Price | Continuous | The price per unit (decimal values allowed) |
| Product | Nominal | The product sold (categorical, no order) |
| CustomerTier | Ordinal | Customer tier: Low, Medium, High (ordered) |

```
# Create the example dataset
sales_data <- data.frame(
  Date = as.Date(c('2025-10-01', '2025-10-01', '2025-10-02', '2025-10-02')),
  Qty = c(2, 5, 1, 3),                                          # Discrete
  Price = c(1000, 20, 1000, 30),                                # Continuous
  Product = c('Laptop', 'Mouse', 'Laptop', 'Keyboard'),         # Nominal
  CustomerTier = factor(c('High', 'Medium', 'Low', 'Medium'), # Ordinal
                        levels = c('Low', 'Medium', 'High'),
                        ordered = TRUE))

print(sales_data)      # View the dataset / str(sales_data).
```

### 2.5.2 Variables (Columns)

**Variables** are the columns or attributes in a dataset that store specific pieces of information about each observation. They define **what kind of data is collected** and determine the types of analysis that can be performed [19].

### 2.5.3 Observations (Rows)

**Observations** are the rows in a dataset, with each row representing a single case, event, or unit of analysis [19]. Together, variables and observations form the core structure of a dataset, allowing us to organize, explore, and analyze data effectively [19].

## References

# Chapter 3

# Basic Data Visualizations

Data visualization is a crucial process in transforming raw data into clear, meaningful, and actionable insights. Before creating effective charts or graphs, it is essential to develop a comprehensive understanding of the data's characteristics, including its type, structure, and key attributes. This foundational understanding ensures that visualizations accurately represent the data and effectively communicate the intended message, thereby minimizing the risk of misinterpretation [21].

**Watch here:** Data Visualization and Misrepresentation

This section focuses on Basic Data Visualizations (Figure 3.1), explaining how data can be categorized into numeric (quantitative) and categorical (qualitative) forms, along with subtypes like discrete, continuous, nominal, and ordinal. It also discusses common data sources and the fundamental elements of a dataset, such as variables and observations, which are essential for selecting appropriate visualization methods.

As discussed in the section of Data Exploration, understanding data types and structure is essential before creating visualizations. By considering the structure of datasets including variables, observations, and data sources readers can select appropriate visual representations, such as histograms for continuous data, bar charts for categorical data, or scatter plots for examining relationships. This thoughtful selection of visualization methods helps reveal patterns, trends, and actionable insights within the dataset [22]; [23].

According to the mindmap, the following section will explore several fundamental data visualizations by emphasizing their types, purposes, applications, users, and tools. Starting with these essential visualizations is crucial before progressing to more advanced analytical techniques. These visuals not only help us understand distributions, comparisons, and relationships between variables in a simple yet informative way but also provide the foundation for deeper analysis. By mastering these basics, we can communicate insights more effectively, spot hidden patterns, and make data-driven decisions with greater confidence [24]–[26]. Before moving forward to next sections, please consider to watching this video.

**Watch here:** Science of Data Visualization

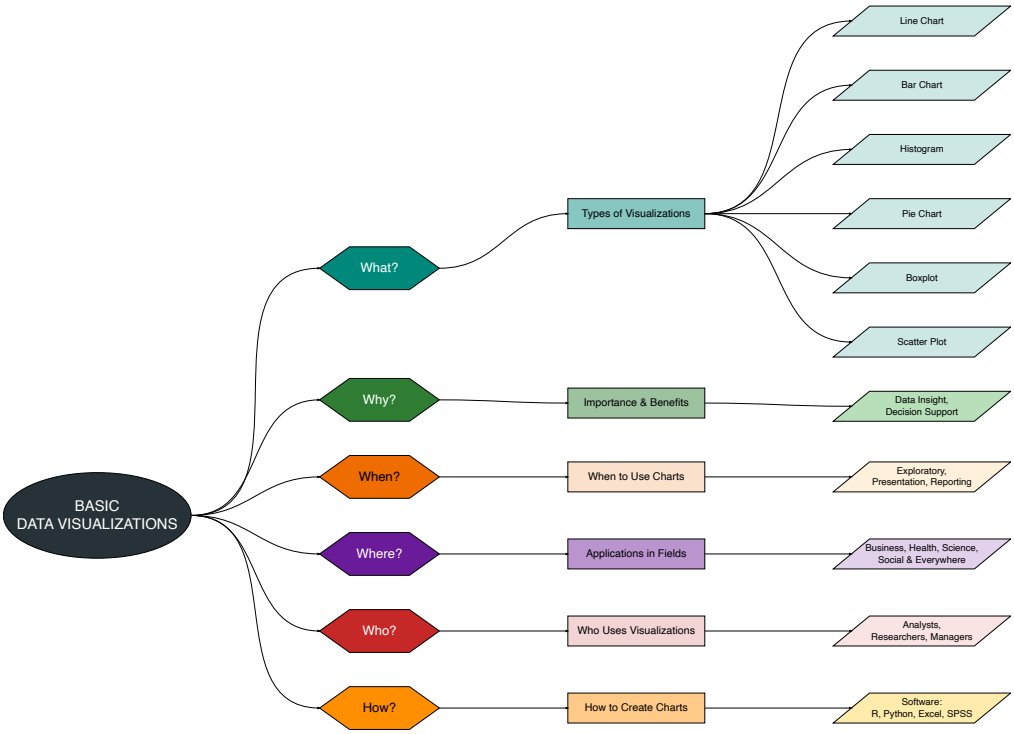



Figure 3.1: Basic Data Visualizations 5W+1H

## 3.1 Dataset

This dataset represents **200 simulated sales transactions** from various cities across Indonesia during the year **2024**. It is designed to illustrate different types of data commonly found in business and analytics contexts — including **nominal**, **ordinal**, **discrete**, and **continuous** variables.

Each row in the dataset corresponds to a single customer transaction, recording essential details such as **date, product type, city, customer tier, quantity sold, price, and payment method**. The dataset is intentionally structured to be used for teaching and practicing **data exploration, visualization, and analysis** in tools like **R, Python, Excel, or Power BI**.

### 3.1.1 Purpose of the Dataset

The dataset can be used to:

- Demonstrate how to identify and classify different data types (nominal, ordinal, discrete, continuous).
- Practice generating and interpreting common visualizations such as **line chart**, **bar charts**, **histograms**, **pie charts**, **boxplots**, and **scatter plots**.
- Perform exploratory data analysis (EDA) on sales trends, customer segments, and pricing patterns.
- Explore relationships between variables, such as how **quantity** and **price** affect total sales or how **customer tiers** differ across **payment methods**.

### 3.1.2 Dataset Overview

| Column | Example | Data Type | Description |
|---|---|---|---|
| `TransactionID` | T0045 | Nominal | Unique identifier for each transaction |
| `TransactionDate` | 2024-05-14 | Date | Date of transaction |
| `ProductCategory` | Electronics | Nominal | Category of the purchased product |
| `City` | Jakarta | Nominal | City where the transaction occurred |
| `CustomerTier` | Gold | Ordinal | Customer level (Bronze < Silver < Gold < Platinum) |
| `Quantity` | 3 | Discrete | Number of items sold |
| `UnitPrice` | 1,200,000 | Continuous | Price per unit of the product |
| `TotalPrice` | 3,600,000 | Continuous | Total transaction value |
| `Advertising` | 500,000 | Continuous | Advertising spend associated with the transaction |

| Column | Example | Data Type | Description |
|---|---|---|---|
| PaymentMethod | Credit Card | Nominal | Payment method used by the customer |

```r
library(DT)
# Generate Sales Transaction Dataset in R
# ==============================

set.seed(123)  # reproducible

# --- 1. Define base variables ---
TransactionID <- sprintf("T%04d", 1:200)

TransactionDate <- sample(seq(as.Date("2025-01-01"), as.Date("2025-12-31"),
                              by = "day"), 200, replace = TRUE)

ProductCategory <- sample(c("Electronics", "Groceries", "Fashion",
                            "Furniture", "Beauty"), 200, replace = TRUE)

City <- sample(c(
  "Jakarta", "Surabaya", "Bandung", "Medan", "Semarang", "Palembang",
  "Makassar", "Bekasi", "Tangerang", "Depok", "Batam", "Pekanbaru",
  "Bandar Lampung", "Denpasar", "Padang", "Malang", "Banjarmasin",
  "Pontianak", "Manado", "Balikpapan"
), 200, replace = TRUE)

CustomerTier <- sample(c("Bronze", "Silver", "Gold", "Platinum"), 200,
                       replace = TRUE, prob = c(0.3, 0.4, 0.2, 0.1))

Quantity <- sample(1:10, 200, replace = TRUE)

UnitPrice <- round(runif(200, 20000, 3000000), 0)

# --- Advertising spend (continuous) ---
Advertising <- round(runif(200, 50000, 1000000), 0)

# --- TotalPrice: positive linear relationship with Advertising ---
# Formula: TotalPrice = base + slope * Advertising + random noise
TotalPrice <- round(50000 + 2 * Advertising +
                      rnorm(200, mean = 0, sd = 50000), 0)

PaymentMethod <- sample(c("Cash", "Credit Card", "Debit Card", "E-Wallet"),
                        200, replace = TRUE)

# --- Combine into a data frame ---
sales_data <- data.frame(
  TransactionID,
  TransactionDate,
```

```
  ProductCategory,
  City,
  CustomerTier,
  Quantity,
  UnitPrice,
  TotalPrice,
  Advertising,
  PaymentMethod
)

# Display the data frame
library(DT)
datatable(sales_data,
          caption = "Dataset with Positive Linear TotalPrice vs Advertising",
          rownames = FALSE)
```

Show 10 entries Search:

Dataset with Positive Linear TotalPrice vs Advertising

| TransactionID | TransactionDate | ProductCategory | City | CustomerTier | Quantity | UnitPrice | TotalPrice | Advertising | PaymentMethod |
|---|---|---|---|---|---|---|---|---|---|
| T0001 | 2025-06-28 | Furniture | Padang | Bronze | 2 | 2172749 | 1177626 | 579427 | Credit Card |
| T0002 | 2025-01-14 | Groceries | Balikpapan | Gold | 2 | 2767574 | 1579042 | 743873 | E-Wallet |
| T0003 | 2025-07-14 | Electronics | Banjarmasin | Gold | 2 | 1849246 | 1040304 | 496366 | Cash |
| T0004 | 2025-11-02 | Groceries | Makassar | Silver | 3 | 499680 | 2043426 | 989180 | Debit Card |
| T0005 | 2025-04-28 | Furniture | Padang | Silver | 8 | 565163 | 2033888 | 985435 | E-Wallet |
| T0006 | 2025-10-26 | Beauty | Semarang | Silver | 9 | 2195947 | 1026012 | 423620 | E-Wallet |
| T0007 | 2025-08-17 | Electronics | Denpasar | Gold | 1 | 1776786 | 1247546 | 628405 | Debit Card |
| T0008 | 2025-09-01 | Electronics | Padang | Platinum | 7 | 951197 | 811466 | 378210 | Debit Card |
| T0009 | 2025-01-14 | Electronics | Padang | Bronze | 8 | 2322927 | 891515 | 465257 | Credit Card |
| T0010 | 2025-06-02 | Beauty | Malang | Gold | 1 | 1022169 | 614842 | 267675 | E-Wallet |

Showing 1 to 10 of 200 entries Previous 1 2 3 4 5 … 20 Next

## 3.2 Line-chart

A **Line Chart** is a data visualization tool that illustrates how values change over a sequence, typically over time. It connects data points with a continuous line, making it ideal for displaying trends and patterns in time-series data [27]. Line charts are particularly useful for:

- **Identifying Seasonal Patterns**: Recognizing recurring fluctuations at regular intervals, such as increased sales during holidays [28].
- **Detecting Growth or Decline Trends**: Observing upward or downward movements in data over time [29].
- **Spotting Peaks or Dips**: Highlighting significant increases or decreases in activity, such as sales spikes during promotions [30].

In this Dataset, we can use a line chart to show how **total sales** or the **number of transactions** change across **dates** during the year 2024.

### 3.2.1 Basic Line-chart

The following line chart using Base R functions (see Figure 3.2) shows the **monthly sales trend** derived from `sales_data`. This visualization helps identify growth patterns, seasonal fluctuations, and overall performance across time periods.

```r
# Step 1: Ensure TransactionDate is in Date format
sales_data$TransactionDate <- as.Date(sales_data$TransactionDate,
                                      format = "%Y-%m-%d")
# Step 2: Calculate total sales per month
sales_trend <- aggregate(TotalPrice ~ format(sales_data$TransactionDate,
                                             "%Y-%m"),
                         data = sales_data, sum)
# Step 3: Rename columns for better clarity
names(sales_trend) <- c("MonthStr", "TotalSales")
# Step 4: Add "-01" to create a complete date format
sales_trend$Month <- as.Date(paste0(sales_trend$MonthStr, "-01"),
                             format = "%Y-%m-%d")
# Step 5: Plot the line chart
plot(
  sales_trend$Month,
  sales_trend$TotalSales,
  type = "o",
  col = "steelblue",
  pch = 16,
  lwd = 2,
  main = "Monthly Sales Trend in 2024",
  xlab = "Month",
  ylab = "Total Sales (IDR)"
)
grid(col = "gray80", lty = "dotted")
```

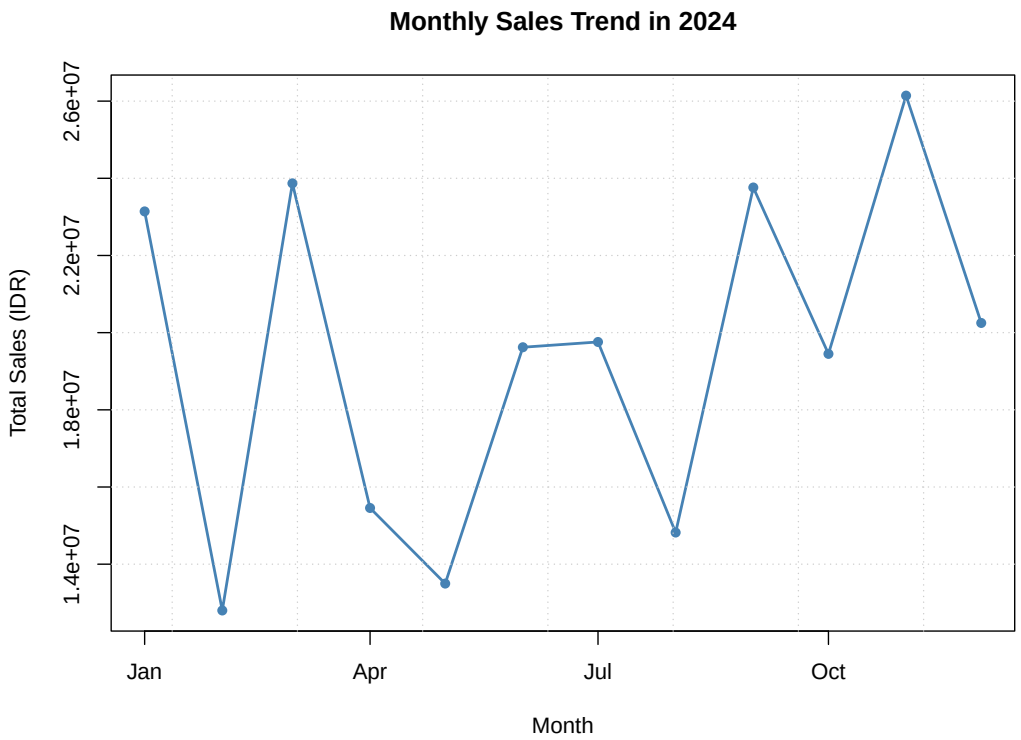



Figure 3.2: Monthly Sales Trend

### 3.2.2 Line-chart using `ggplot2`

The following visualization (Figure 3.3) displays the trend of monthly total sales throughout the year. It helps identify periods of high or low sales performance, supporting time-based decision-making. We use the `ggplot2` library for cleaner visualization and `dplyr + lubridate` for data wrangling.

```
# Load required packages
library(ggplot2)
library(dplyr)
library(lubridate)

# Summarize total sales by month
sales_trend <- sales_data %>%
  mutate(Month = floor_date(TransactionDate, "month")) %>%
  group_by(Month) %>%
  summarise(TotalSales = sum(TotalPrice))

# Create line chart
ggplot(sales_trend, aes(x = Month, y = TotalSales)) +
  geom_line(color = "steelblue", linewidth = 1.2) +  # updated aesthetic
  geom_point(color = "darkorange", size = 2) +
  labs(
```

```
    title = "Monthly Sales Trend in 2024",
    x = "Month",
    y = "Total Sales (IDR)"
  ) +
  theme_minimal()
```

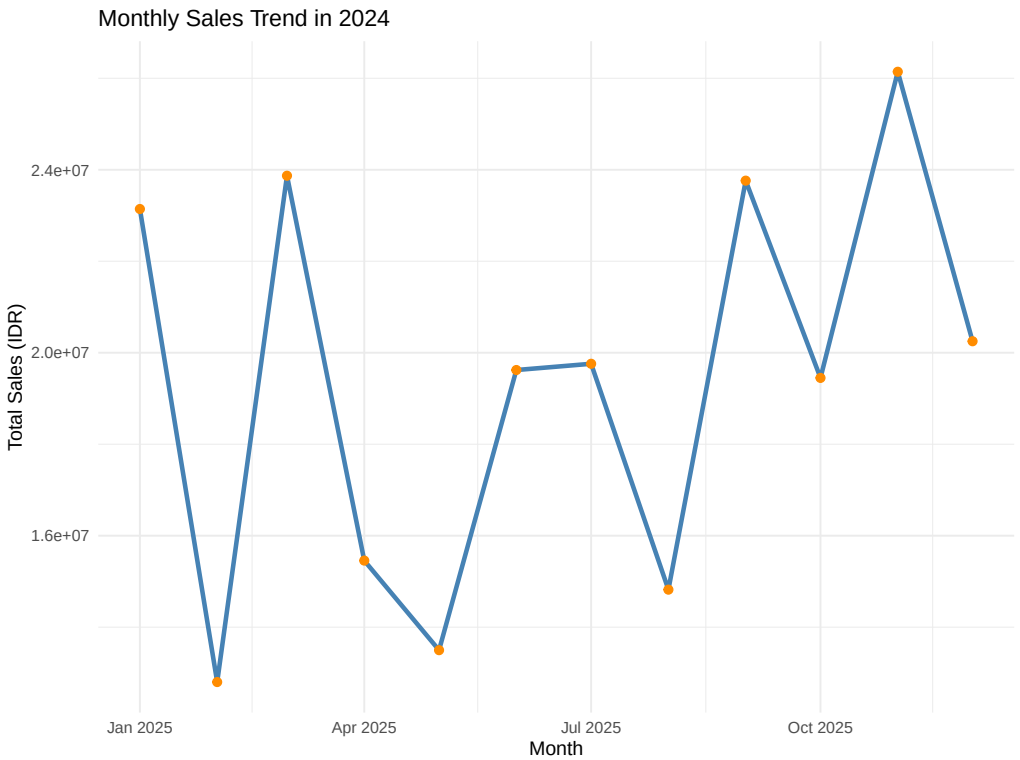



Figure 3.3: Monthly Sales Trend (ggplot2)

## 3.3 Bar-chart

A **Bar Chart** is a type of data visualization used to represent **categorical data** with rectangular bars. Each bar's height (or length) corresponds to the value or frequency of a category, making it easy to compare quantities across different groups [31].

Bar charts are **especially suitable for**:

- **Discrete numeric data** – numbers that can only take specific values (e.g., number of items purchased) [32].
- **Ordinal categorical data** – categories with a natural order (e.g., customer satisfaction levels: Low, Medium, High) [33].

In this Dataset, the **Bar Chart** is used to show the **Total Sales by City**. This allows us to quickly identify which cities contribute the most to total sales performance [34].

**Insights:**

- Taller bars indicate higher total sales.
- The chart helps compare city-level sales performance visually.
- It is ideal for **categorical variables** such as `City` and **discrete numeric values** like `TotalPrice`.
- For ordinal data, bar charts make it easy to observe trends or patterns across ordered categories.

### 3.3.1 Basic Bar-chart

In this section, we create a bar chart (see, Figure 3.4) using Base R functions instead of ggplot2. The base plotting system in R provides a simple and direct way to visualize data without requiring additional packages. Here, we visualize total sales by city to compare which locations contribute most to overall revenue.

```r
# Step 1: Aggregate total sales per city
sales_city <- aggregate(TotalPrice ~ City, data = sales_data, sum)

# Step 2: Sort data by total sales (descending)
sales_city <- sales_city[order(sales_city$TotalPrice, decreasing = TRUE), ]

# Step 3: Set margins
par(mar = c(8, 5, 4, 2))  # c(bottom, left, top, right)

# Step 4: Create bar chart
barplot(
  height = sales_city$TotalPrice,
  names.arg = sales_city$City,
  col = "steelblue",
  las = 2,                    # rotate city labels vertically
  cex.names = 0.8,            # reduce font size of city names
  main = "Total Sales by City",
  xlab = "",
  ylab = ""
)

# Optional: Add grid lines
grid(nx = NA, ny = NULL, col = "gray80", lty = "dotted")
```

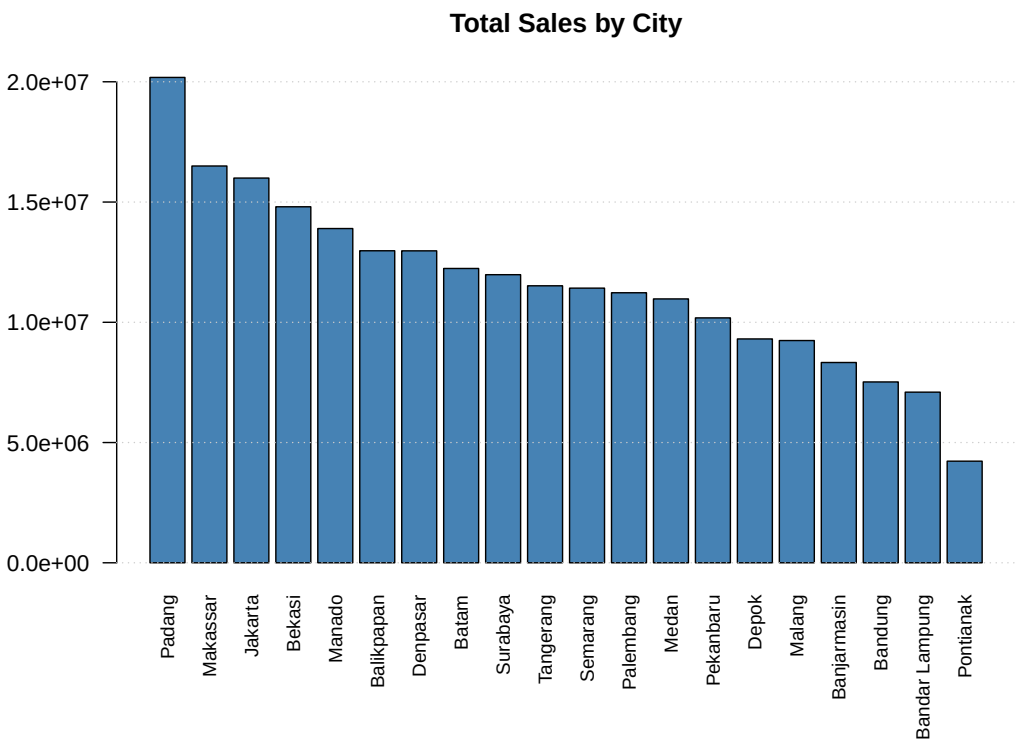



Figure 3.4: Total Sales by City

### 3.3.2 Bar-chart using `ggplot2`

In this section, we create the same bar chart (see, Figure 3.5) using the `ggplot2` package, which provides a more modern and flexible approach to visualization in R. Compared to the Base R plotting system, `ggplot2` allows easier customization, better control over aesthetics, and integration with themes and color palettes. We visualize total sales by city to compare sales performance across locations.

```
# Load ggplot2
library(ggplot2)

# Summarize total sales per city
sales_city <- aggregate(TotalPrice ~ City, data = sales_data, sum)

# Sort city by total sales (descending)
sales_city <- sales_city[order(sales_city$TotalPrice, decreasing = TRUE), ]

# Create bar chart
ggplot(sales_city, aes(x = reorder(City, -TotalPrice), y = TotalPrice)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(TotalPrice/1e6, 1)),
            vjust = -0.5, size = 3, color = "black") +
  labs(
```

```
    title = "Total Sales by City",
    x = "City",
    y = "Total Sales (in Millions IDR)"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 9),
    plot.title = element_text(size = 14, face = "bold")
  )
```

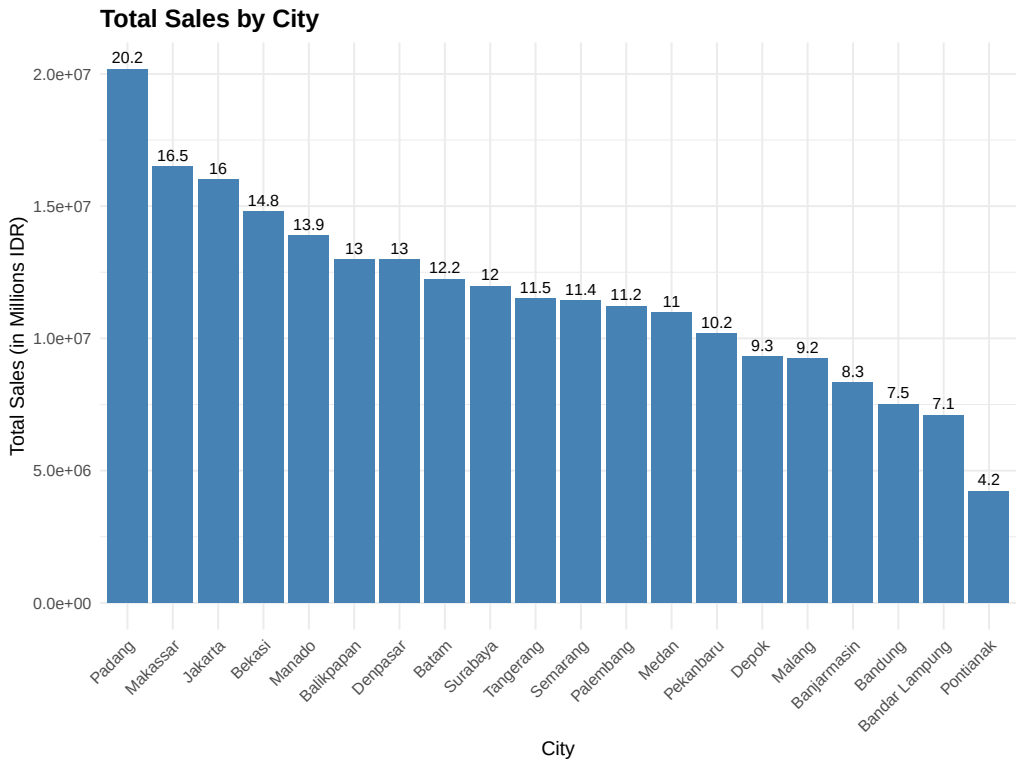



Figure 3.5: Total Sales by City (ggplot2)

## 3.4 Histogram-chart

A **Histogram** is a graphical representation of the distribution of numerical data. It divides the data into intervals, known as bins, and displays the frequency of data points within each bin. This visualization helps identify patterns such as the central tendency, spread, skewness, and the presence of multiple modes in the data [27].

Histograms are particularly effective for:

- **Visualizing the Distribution**: They provide a clear picture of how data is distributed across different ranges, helping to identify the shape of the distribution (e.g., normal,

skewed, bimodal) [34].

- **Identifying Central Tendency and Spread**: By observing the peak of the histogram, one can infer the central value of the data. The width of the histogram indicates the variability or spread of the data [35].
- **Detecting Skewness**: The asymmetry of the histogram can reveal whether the data is skewed to the left or right, indicating potential biases in the data collection process [36].
- **Recognizing Multiple Modes**: A histogram can show if the data has multiple peaks (modes), suggesting the presence of different subgroups within the dataset [37].

In this Dataset, we can use histograms to explore the distribution of variables such as:

- `Quantity` (number of items purchased)
- `UnitPrice` (price per item)
- `TotalPrice` (total transaction value)

### 3.4.1 Basic Histogram-chart

In this example, we use Base R plotting functions to create a histogram (Figure 3.6) showing how many transactions occurred for each quantity of items purchased. A histogram helps us understand the distribution of data — in this case, which purchase quantities are most common. Peaks (tall bars) represent quantities that occur more frequently, while shorter bars indicate rarer purchase sizes.

```
hist(sales_data$Quantity,
     main = "Histogram of Quantity",
     xlab = "Number of Items Purchased",
     ylab = "Frequency",
     col = "skyblue",
     border = "white",
     breaks = 5)
```

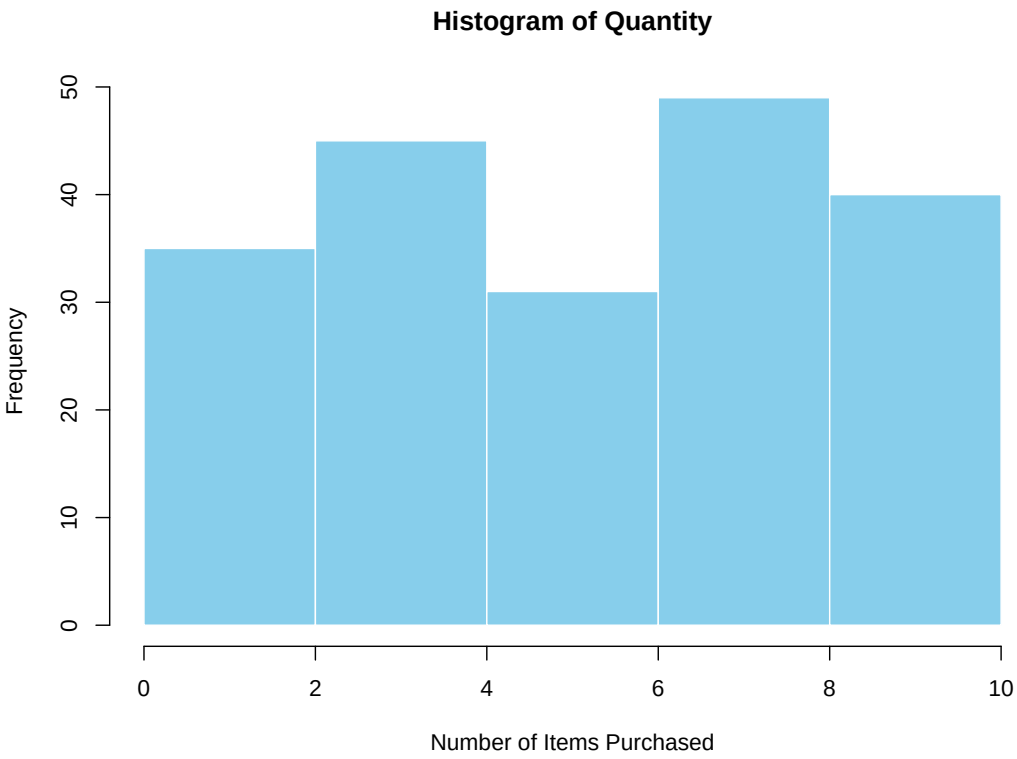



Figure 3.6: Histogram of Quantity

### 3.4.2 Histogram-chart using `ggplot2`

In this case, we use the `ggplot2` library to display how many transactions fall within each range of item quantities purchased. Each bar represents a range of purchase quantities, while the height indicates the frequency of transactions within that range. This Figure 3.7 helps us quickly identify whether customers tend to buy in small, medium, or large quantities.

```
library(ggplot2)

ggplot(sales_data, aes(x = Quantity)) +
  geom_histogram(
    bins = 5,                # number of bins (adjust as needed)
    fill = "skyblue",        # fill color for the bars
    color = "white",         # border color for the bars
    alpha = 0.8              # transparency level
  ) +
  labs(
    title = "Histogram of Quantity",
    x = "Number of Items Purchased",
    y = "Frequency"
  ) +
  theme_minimal()
```

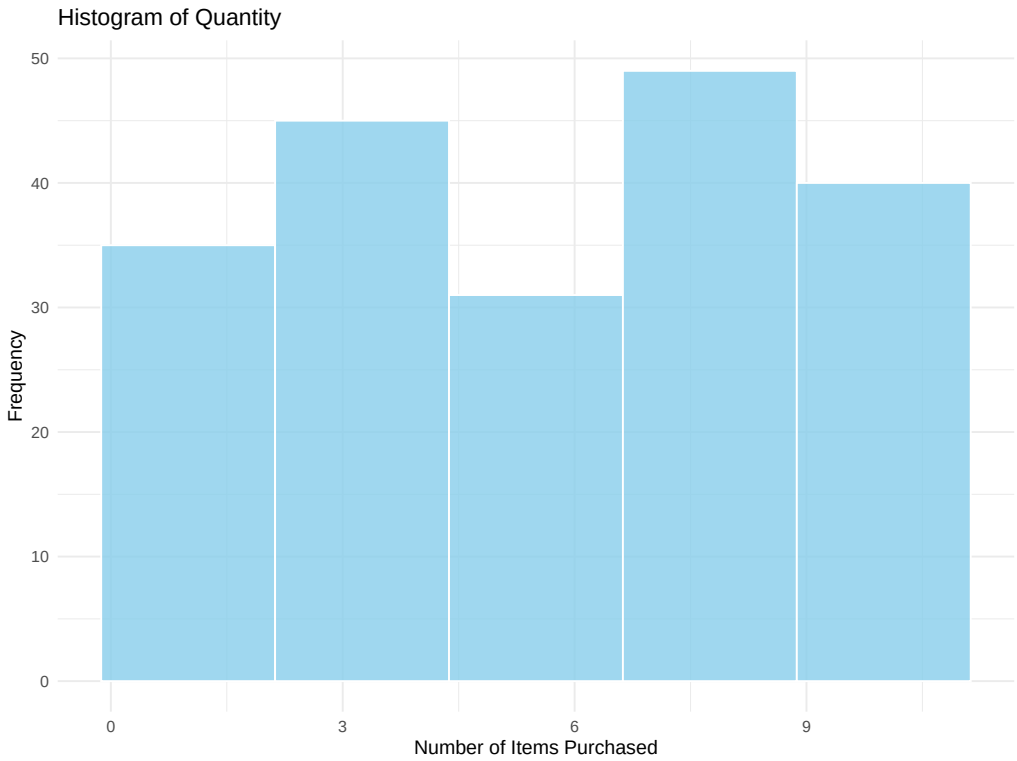



Figure 3.7: Histogram of Quantity (ggplot2)

## 3.5 Pie-chart

A **Pie Chart** is a circular statistical graphic divided into slices to illustrate **numerical proportions** within a dataset. Each slice of the pie represents a category's contribution to the **whole**, making it ideal for showing **part-to-whole relationships**.

Pie charts are best used when:

- The dataset contains a **small number of categories**.
- You want to emphasize **relative proportions** or **percentages**.
- The total adds up to **100%** of the dataset.

However, pie charts are **less effective** when there are too many categories or when the differences between slices are small — in such cases, a **bar chart** is often more suitable [38]; [39].

### 3.5.1 Basic Pie-chart

In this dataset, we can use a **Pie Chart** (see, Figure 3.8) to visualize the **percentage contribution of total sales by product category**. This helps to quickly understand which product categories dominate total sales performance.

```r
# --- Summarize total sales by product category (base R only) ---
total_sales <- tapply(sales_data$TotalPrice, sales_data$ProductCategory, sum)

# Calculate percentage for each category
percentage <- round(100 * total_sales / sum(total_sales), 1)

# Create labels with category names and percentage
labels <- paste(names(total_sales), " - ", percentage, "%", sep = "")

# --- Create Donut Chart (Base R) ---
par(mar = c(2, 2, 2, 2))  # Adjust margins for clean layout

# Draw pie chart
pie(
  total_sales,
  labels = labels,
  main = "Percentage of Total Sales by Product Category (2024)",
  col = rainbow(length(total_sales)),
  clockwise = TRUE,
  border = "white",
  radius = 1,
  cex = 0.9     # control label size
)

# Add a white circle in the center to make it a donut
symbols(
  0, 0,
  circles = 0.4,
  inches = FALSE,
  add = TRUE,
  bg = "white",  # center color (donut hole)
  fg = NA        # remove border
)
```

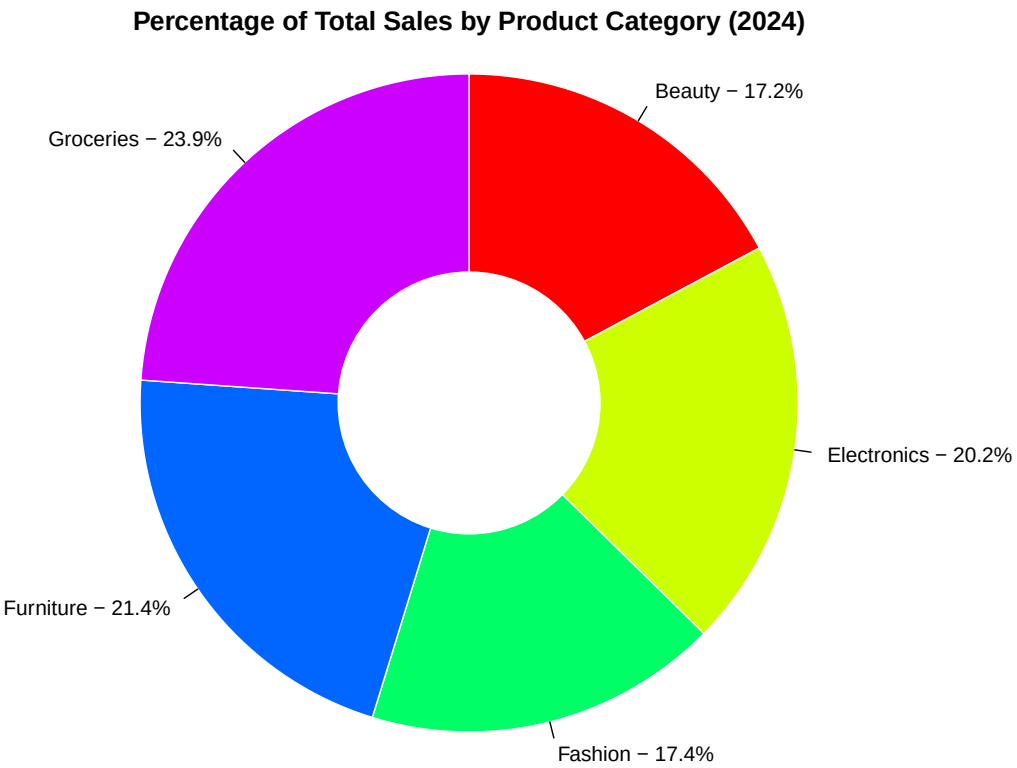



Figure 3.8: Percentage of Total Sales by Product Category (2024)

**Insights:**

- Larger slices indicate higher total sales share.
- Useful for summarizing **categorical variables** such as `ProductCategory`.
- Supports decision-making by highlighting the dominant categories in the market.

### 3.5.2 Pie-chart using `ggplot2`

In this example, we use the ggplot2 package to visualize the percentage of total sales by city (see, Figure 3.9). Each slice of the pie represents one city's contribution, making it easy to compare which cities generate the most or least revenue.

```
library(ggplot2)
library(dplyr)
library(ggrepel)

# --- Summarize total sales by product category ---
sales_summary <- sales_data %>%
  group_by(ProductCategory) %>%
```

```
  summarise(TotalSales = sum(TotalPrice, na.rm = TRUE)) %>%
  mutate(Percentage = round(100 * TotalSales / sum(TotalSales), 1)) %>%
  arrange(desc(TotalSales)) %>%
  mutate(
    ypos = cumsum(TotalSales) - 0.5 * TotalSales,
    Label = paste0(ProductCategory, "\n", Percentage, "%")
  )

# --- Donut Chart (Polished Version) ---
ggplot(sales_summary, aes(x = 2, y = TotalSales, fill = ProductCategory)) +
  geom_col(width = 1, color = "white") +
  coord_polar(theta = "y", start = 0) +
  xlim(0.5, 3) +   # add extra space around the donut
  theme_void() +
  geom_text_repel(
    aes(y = ypos, label = Label),
    color = "black",
    size = 4,
    nudge_x = 1,                # push labels away from the donut
    force = 8,                  # increase label repulsion strength
    segment.size = 0.5,
    segment.color = "gray60",
    min.segment.length = 0.5,
    max.overlaps = Inf,
    show.legend = FALSE
  ) +
  scale_fill_brewer(palette = "Set2") +
  annotate("text", x = 0.5, y = 0, label = "Total Sales",
           size = 5, color = "gray40") +
  labs(
    title = "Percentage of Total Sales by Product Category (2024)",
    fill = "Product Category"
  ) +
  theme(
    plot.title = element_text(
      hjust = 0.5,
      face = "bold",
      size = 14,
      color = "#333333"
    ),
    legend.position = "none",
    plot.background = element_rect(fill = "white", color = NA)
  )
```

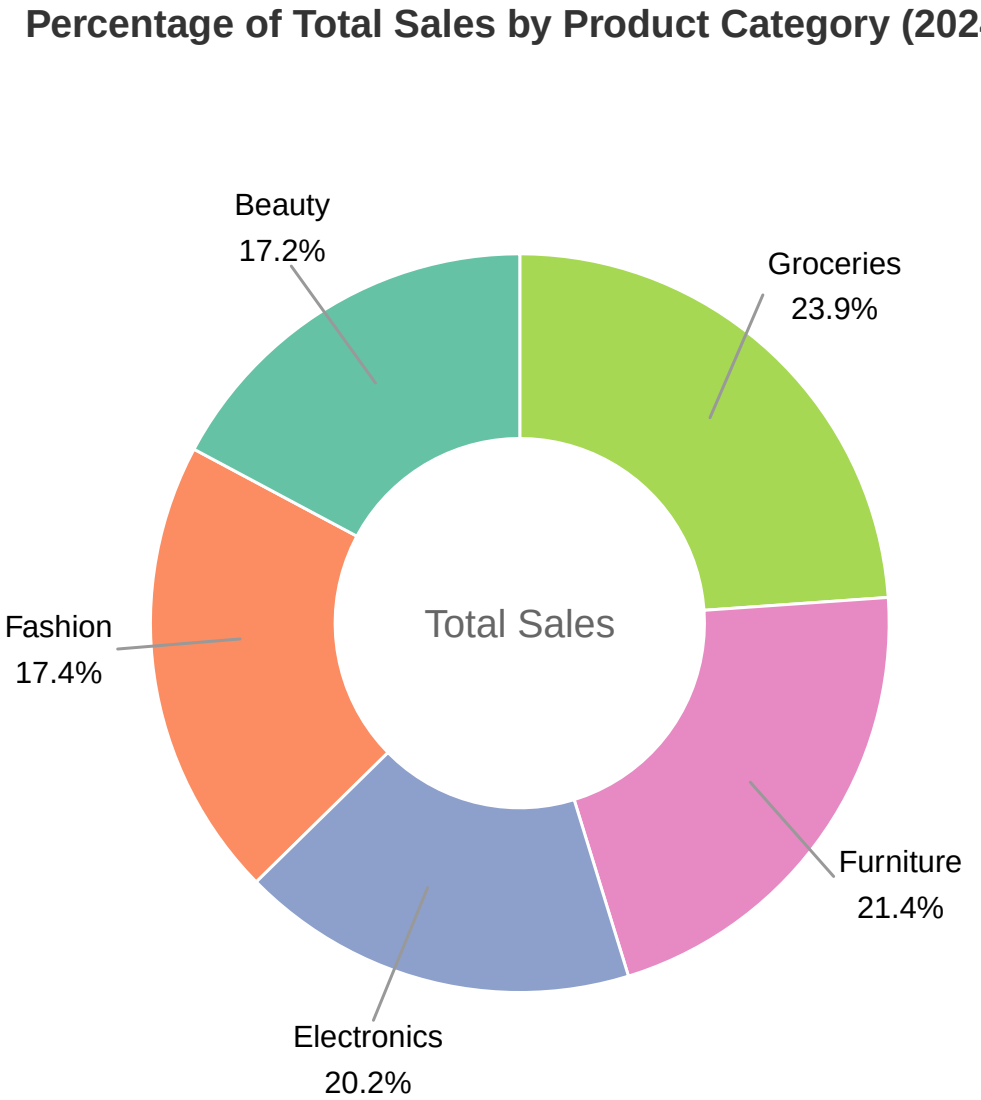



Figure 3.9: Percentage of Total Sales by Product Category (2024) (ggplot2)

## 3.6 Box-plot

A **Boxplot** is a data visualization technique that displays the **distribution, spread, and potential outliers** of a continuous variable through its summary statistics — minimum, first quartile (Q1), median, third quartile (Q3), and maximum [27]. It provides a compact view of how data values are dispersed and where they concentrate.

Boxplots are particularly useful for: - **Comparing Distributions Across Groups**: Revealing differences in data spread and central tendency across categories (e.g., product types or customer tiers) [28]. - **Detecting Outliers**: Identifying unusually high or low data points that may indicate data errors or special cases [29]. - **Assessing Data Symmetry and Skewness**: Observing

whether the data are evenly distributed or skewed toward higher or lower values [30].

In this Dataset, a boxplot can be used to **compare the distribution of total sales (`TotalPrice`) across different customer tiers (`CustomerTier`) or product categories (`ProductCategory`)**. This helps identify which segments tend to have higher transaction values and whether there are significant outliers in purchasing behavior.

### 3.6.1 Basic Box-plot

In this example, we use Base R plotting to display how total sales (TotalPrice) vary across different product categories (see,Figure 3.10). The box represents the interquartile range (IQR), the line inside shows the median, and dots beyond the whiskers may indicate outliers.

```
# ==========================================
# Boxplot of TotalPrice with arrows & stats
# Base R version (no ggplot2)
# ==========================================

# Compute summary statistics
summary_stats <- data.frame(
  Stat  = c("Min", "Q1", "Median", "Q3", "Max", "Mean"),
  Value = c(
    min(sales_data$TotalPrice, na.rm = TRUE),
    quantile(sales_data$TotalPrice, 0.25, na.rm = TRUE),
    median(sales_data$TotalPrice, na.rm = TRUE),
    quantile(sales_data$TotalPrice, 0.75, na.rm = TRUE),
    max(sales_data$TotalPrice, na.rm = TRUE),
    mean(sales_data$TotalPrice, na.rm = TRUE)
  )
)

# --- Adjust small offset to avoid overlap between Median and Mean ---
median_idx <- which(summary_stats$Stat == "Median")
mean_idx   <- which(summary_stats$Stat == "Mean")

# If values are close, add vertical offset
if (abs(summary_stats$Value[median_idx] - summary_stats$Value[mean_idx]) <
    0.05 * diff(range(summary_stats$Value))) {
  summary_stats$Value[median_idx] <-
    summary_stats$Value[median_idx] * 1.02  # move slightly upward
  summary_stats$Value[mean_idx]   <-
    summary_stats$Value[mean_idx]   * 0.95  # move slightly downward
}

# Adjust plot margins (more space on the right)
par(mar = c(5, 4, 5, 6))

# Create basic boxplot
boxplot(
```

```
  sales_data$TotalPrice,
  main = "Distribution of Total Sales Value with Summary Statistics",
  ylab = "Total Price (IDR)",
  col = "#69b3a2",
  border = "gray40",
  boxwex = 0.3,
  notch = FALSE,
  outline = TRUE
)

# X position of the boxplot
x_box <- 1

# Add arrows pointing from left to right
arrows(
  x0 = x_box + 0.15, x1 = x_box + 0.25,
  y0 = summary_stats$Value, y1 = summary_stats$Value,
  length = 0.08, angle = 20, col = "gray40", lwd = 1
)

# Add text labels to the right of the arrows
text(
  x = x_box + 0.28, y = summary_stats$Value,
  labels = paste0(summary_stats$Stat, ": ",
                  format(round(summary_stats$Value, 0), big.mark = ",")),
  pos = 4,  # left-aligned text
  cex = 0.8, col = "#222222"
)

# Add caption below the plot
mtext("Source: @dsciencelabs",
      side = 1,
      line = 4,
      adj = 1,
      cex = 0.8,
      col = "gray50")
```

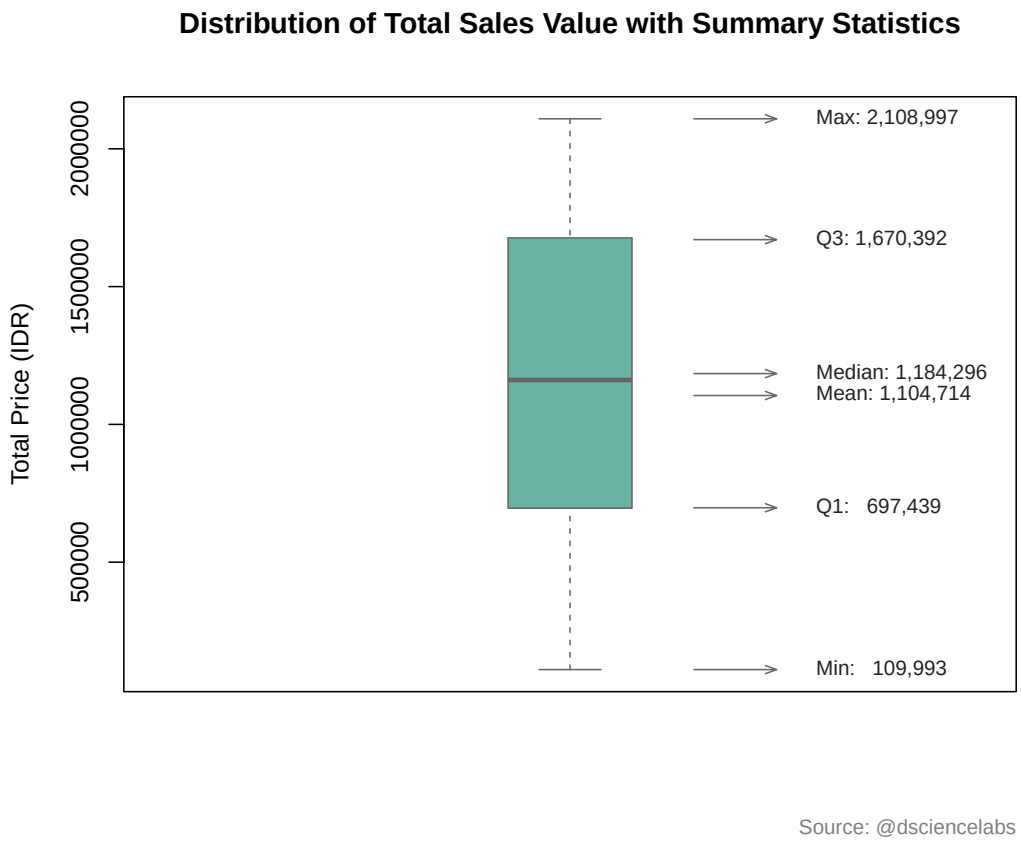



Figure 3.10: Distribution of Total Sales Value with Summary Statistics

**Explanation:**

| Statistic | Description |
|---|---|
| **Min** | The smallest value in the dataset. |
| **Q1 (First Quartile)** | The value below which 25% of the data fall. |
| **Median** | The middle value dividing the data into two equal halves. |
| **Q3 (Third Quartile)** | The value below which 75% of the data fall. |
| **Max** | The largest value in the dataset. |
| **Mean** | The arithmetic average of all data points. |

### 3.6.2 Box-plot using `ggplot2`

The ggplot2 package allows a more elegant and customizable approach to creating box plots (see, Figure 3.11). This visualization shows the distribution of total sales (TotalPrice) across product categories, highlighting the median, variability, and potential outliers. Compared to the Base R version, ggplot2 provides smoother visuals and easier styling through themes and color mapping.

```r
library(ggplot2)
library(dplyr)
library(grid)

# ============================================================
# Compute summary statistics for TotalPrice
# ============================================================
summary_stats <- data.frame(
  Stat  = c("Min", "Q1", "Median", "Q3", "Max", "Mean"),
  Value = c(
    min(sales_data$TotalPrice, na.rm = TRUE),
    quantile(sales_data$TotalPrice, 0.25, na.rm = TRUE),
    median(sales_data$TotalPrice, na.rm = TRUE),
    quantile(sales_data$TotalPrice, 0.75, na.rm = TRUE),
    max(sales_data$TotalPrice, na.rm = TRUE),
    mean(sales_data$TotalPrice, na.rm = TRUE)
  )
)

# ============================================================
# Adjust small offset to avoid overlap between Median & Mean
# ============================================================
median_idx <- which(summary_stats$Stat == "Median")
mean_idx   <- which(summary_stats$Stat == "Mean")

if (abs(summary_stats$Value[median_idx] - summary_stats$Value[mean_idx]) <
    0.05 * diff(range(summary_stats$Value))) {
  summary_stats$Value[median_idx]<-summary_stats$Value[median_idx] * 1.02 # up
  summary_stats$Value[mean_idx] <- summary_stats$Value[mean_idx]   * 0.98 # down
}

# ============================================================
# Create ggplot boxplot with larger width and arrows
# ============================================================
ggplot(sales_data, aes(x = "Total", y = TotalPrice)) +
  geom_boxplot(
    width = 0.6,
    fill = "#69b3a2",
    color = "gray40",
    outlier.color = "gray40"
  ) +

  # Horizontal arrows shifted right to avoid overlapping the boxplot
  geom_segment(
    data = summary_stats,
    aes(x = 1.65, xend = 1.8, y = Value, yend = Value),
    arrow = arrow(length = unit(0.15, "cm"), angle = 20),
    color = "gray40"
  ) +
```

```
# Text labels at the right of arrows
geom_text(
  data = summary_stats,
  aes(
    x = 1.85,
    y = Value,
    label = paste0(Stat, ": ", format(round(Value, 0), big.mark = ","))
  ),
  hjust = 0,
  size = 3.5,
  color = "#222222"
) +

# Scale and margins to keep everything centered
scale_x_discrete(expand = expansion(add = c(3, 3))) +
labs(
  title = "Distribution of Total Sales Value with Summary Statistics",
  y = "Total Price (IDR)",
  x = NULL,
  caption = "Source: @dsciencelabs"
) +
theme_minimal(base_size = 12) +
theme(
  plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
  plot.caption = element_text(hjust = 0.5, color = "gray50", size = 10),
  axis.text.x = element_blank(),
  axis.ticks.x = element_blank(),
  panel.grid.major.x = element_blank(),
  panel.grid.minor.x = element_blank(),
  plot.margin = margin(20, 120, 20, 120)
) +
coord_cartesian(clip = "off")
```

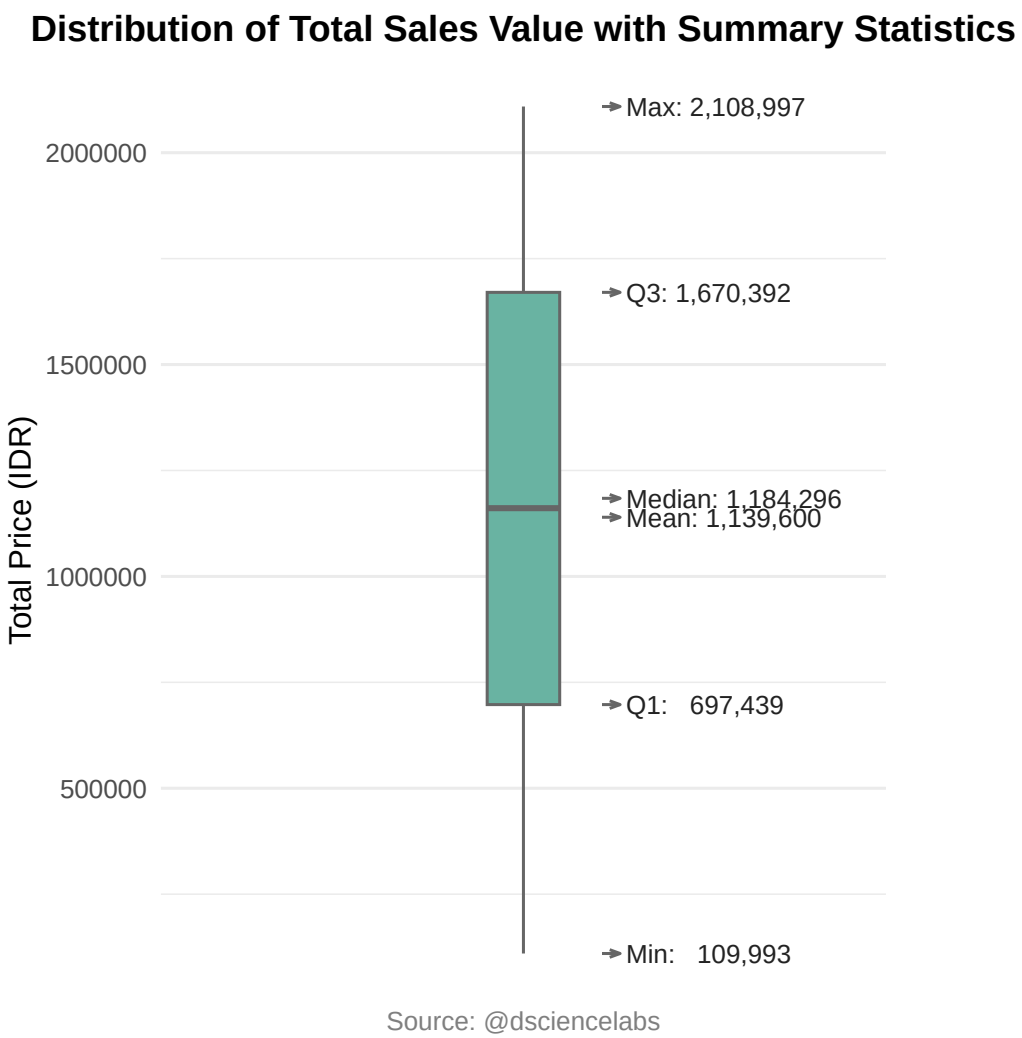



Figure 3.11: Distribution of Total Sales Value with Summary Statistics

The Figure 3.12 provides a visual summary of the spread and central tendency of sales data by Product Category. Each box represents the interquartile range (IQR), the line inside the box marks the median, and points outside the whiskers indicate outliers — unusually high or low sales values that may deserve further investigation.

```
library(ggplot2)

# --- Boxplot of Total Price by Product Category ---

ggplot(sales_data, aes(x = ProductCategory, y = TotalPrice, fill = ProductCategory)) +
geom_boxplot(
color = "darkgray",
outlier.colour = "red",
outlier.shape = 16,
```

```
outlier.size = 2,
alpha = 0.8
) +
labs(
title = "Distribution of Total Sales by Product Category",
x = "Product Category",
y = "Total Price (IDR)",
caption = "Source: @dsciencelabs"
) +
theme_minimal(base_size = 13) +
theme(
plot.title = element_text(
face = "bold",
size = 14,
color = "#333333",
hjust = 0.5
),
axis.text.x = element_text(
angle = 30,
hjust = 1,
color = "#333333"
),
legend.position = "none",
plot.background = element_rect(fill = "white", color = NA)
)
```

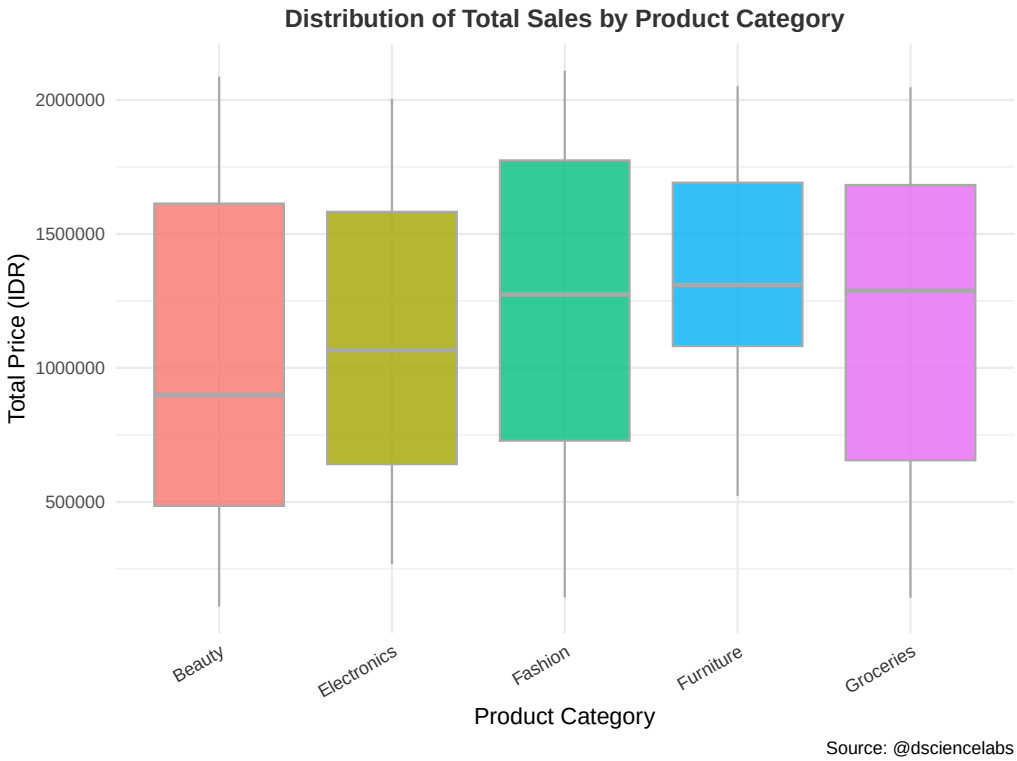



Figure 3.12: Distribution of Total Sales by Product Category

## 3.7 Scatter-plot

A **Scatter Plot** is a data visualization technique that displays the **relationship between two continuous variables** by plotting individual data points on a two-dimensional plane. Each point represents one observation, with its position determined by the values of the two variables. Scatter plots are useful for identifying patterns, trends, correlations, clusters, and potential outliers in the data [27].

Scatter plots are particularly useful for:

- **Identifying Relationships Between Variables**: Revealing positive, negative, or no correlation between variables (e.g., advertising spend vs. sales performance) [28].
- **Detecting Clusters or Groups**: Highlighting natural groupings in data that may correspond to categories, regions, or segments [29].
- **Spotting Outliers**: Identifying unusual data points that deviate from the general trend, which could indicate errors or special cases [30].

In this Dataset, a scatter plot can be used to **explore the relationship between total sales (`TotalPrice`) and advertising spend (`Advertising`)**, or between `TotalPrice` and

`CustomerSatisfaction`. This helps identify whether higher spending leads to increased sales, whether specific groups form distinct clusters, and whether there are extreme observations in the dataset that require further investigation.

### 3.7.1 Basic Scatter-plot

The basic scatter plot shown in Figure 3.13 illustrates the relationship between **Advertising Spend** and **Total Sales (TotalPrice)** using base R plotting. Each point represents a sales observation, allowing us to visually identify patterns or potential outliers. This basic visualization provides a clear starting point before enhancing the design using `ggplot2`.

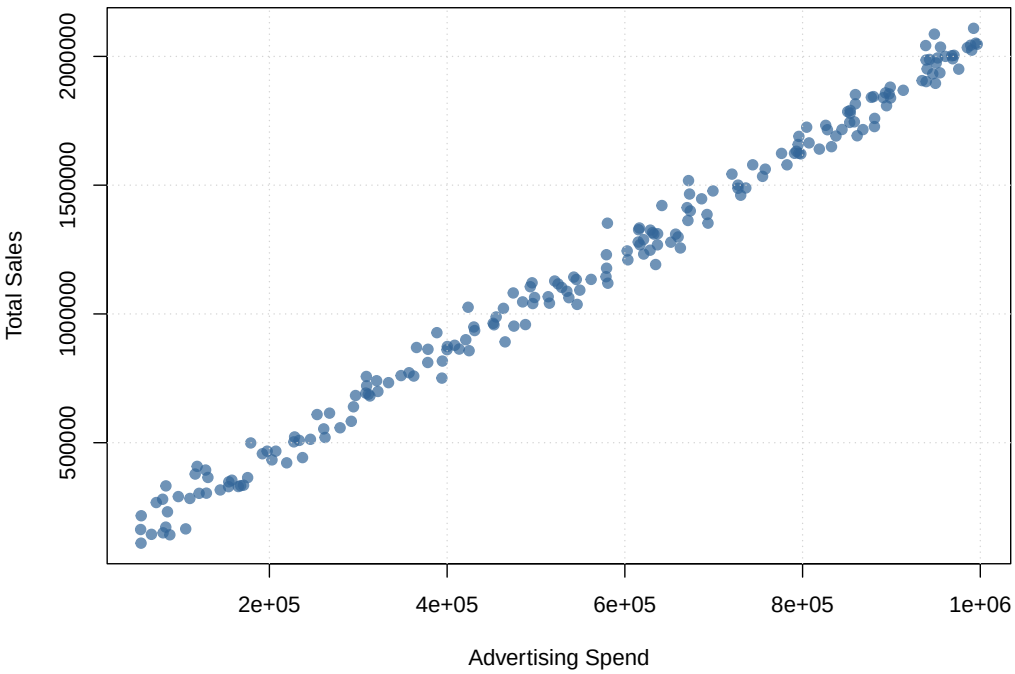



Figure 3.13: Scatter Plot: Total Sales vs Advertising

### 3.7.2 Scatter-plot using `ggplot2`

The Figure 3.14 illustrates the relationship between Advertising Spend and Total Sales (TotalPrice). This visualization helps identify patterns, trends, or potential outliers between these two numerical variables.

## 3.8 Summary

The Table 3.3 provides an overview of the most common chart types used in data visualization, highlighting their advantages, disadvantages, suitable data types, and typical use cases. It serves as a quick reference to help select the most appropriate chart based on the dataset and analytical objectives.

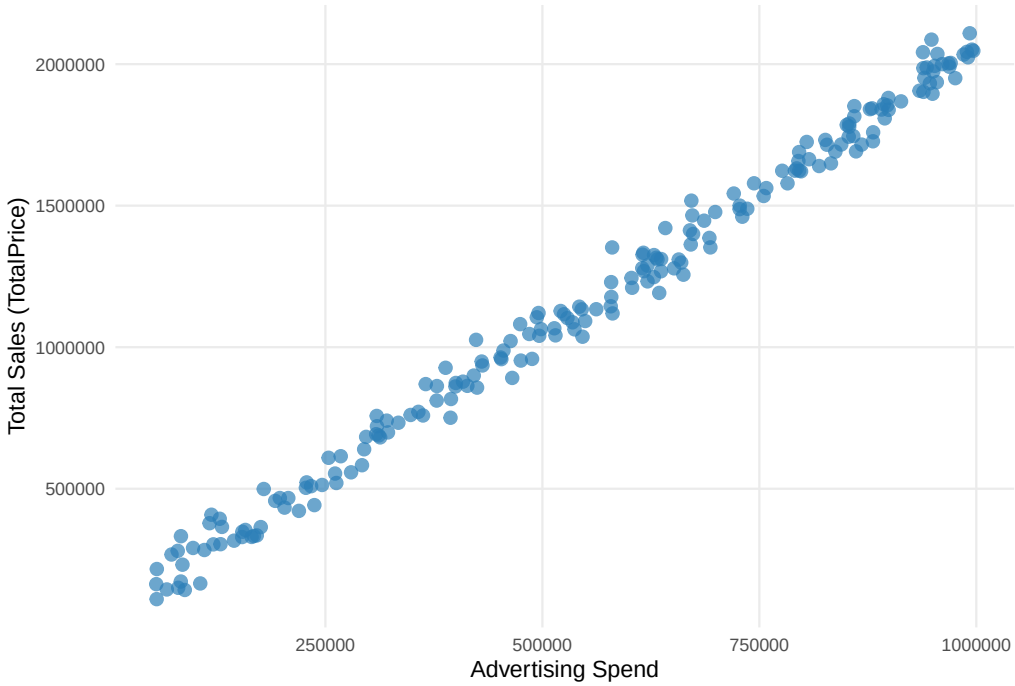



Figure 3.14: Scatter Plot of Total Sales vs Advertising

Table 3.3: Summary of Basic Chart Types

| | Type_Data | Advantages | Disadvantages |
|---|---|---|---|
| Line Chart | Continuous, Time-series | Shows trends over time; easy to read | Not effective for discrete categories; too many lines can be confusing |
| Bar Chart | Categorical, Discrete | Easy comparison between categories; clear visualization | Not suitable for continuous data; crowded if many categories |
| Histogram | Continuous | Shows data distribution; easy to see frequency | Does not show relationship between variables; binning affects interpretation |
| Pie Chart | Categorical | Simple view of proportions or percentages | Hard to compare categories if many; inaccurate for similar values |
| Boxplot | Continuous | Shows distribution, outliers, median, and quartiles | Does not show trend; individual data points are not visible |
| Scatter Plot | Continuous, Numeric | Shows relationship/correlation between two variables; detects outliers | Hard to read if too dense; does not show overall distribution |

# References

# Chapter 4

# Central Tendency

As discussed in the Data Overview section, understanding data types is crucial before applying measures of **Central Tendency (CT)**. For example, the mean is suitable for interval or ratio data, while the median can be applied to both ordinal and continuous data. The mode, however, can be used for all data types, including nominal categories. Choosing the right measure ensures that the "center" of the data is represented accurately, avoiding misleading interpretations.

**Watch here:** Measures of Central Tendency

By mastering central tendency (Figure 4.1), readers will be able to describe datasets more effectively, compare groups of data, and prepare for deeper statistical analysis, such as measures of dispersion and hypothesis testing. Graphical tools—such as histograms, boxplots, and frequency distributions—can further enhance understanding by visually confirming how the data's center aligns with its overall shape and spread [40].

As illustrated in the Figure 4.1, the discussion now turns to measures of central tendency—**mean, median, and mode**—together with guidance on selecting the **most suitable measure** for a given dataset. These statistical tools offer concise summaries of complex information, making it easier to detect patterns, describe distributions, and lay the groundwork for deeper analysis. Gaining proficiency with these measures equips us to interpret data more reliably and to support conclusions with stronger evidence [41], [42].

## 4.1 Definition of CT

**Central Tendency** is a statistical measure that represents the typical or central value of a dataset. It aims to provide **a single value that best represents the entire data**, allowing us to understand where most data values are concentrated. The three most common measures of central tendency are: **Mean**, **Median**, and **Mode** [43].

### 4.1.1 Mean

The **mean** is obtained by dividing the sum of all data values by the total number of observations. It is suitable for **interval** and **ratio** data types.

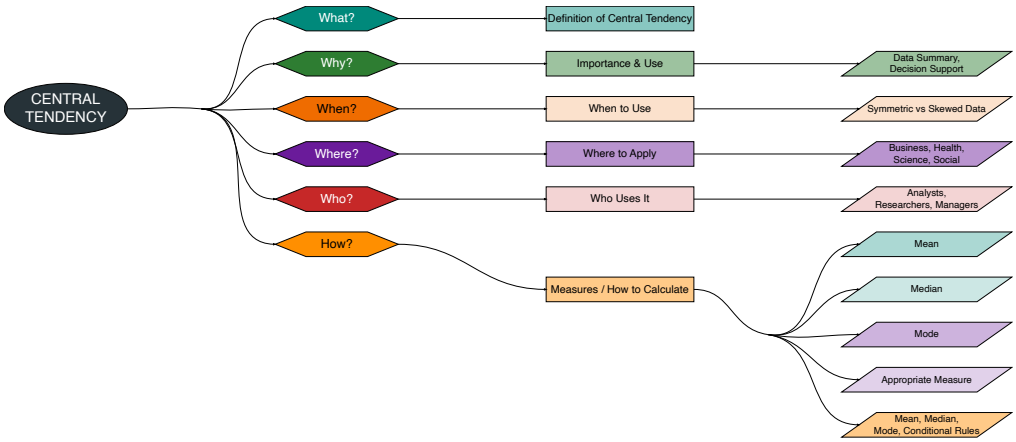



Figure 4.1: Central Tendency 5W+1H

$$\bar{X} = \frac{\sum X_i}{n}$$

**Where:**

- $\bar{X}$: mean (average)
- $X_i$: each data value
- $n$: number of observations

> **Example**
>
> Data: 10, 20, 30, 40, 50
>
> $$\bar{X} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30$$
>
> The average value of the data is **30**.

### 4.1.2 Median

The **median** is the **middle value** of an ordered dataset. It is suitable for **ordinal**, **interval**, and **ratio** data [44]. Steps to Find the Median:

1. Arrange the data in ascending order.
2. If the number of data points $n$ is **odd**, the median is at position $\frac{n+1}{2}$.
3. If $n$ is **even**, the median is the average of the two middle values.

> **Example**
>
> Data: 5, 7, 8, 12, 15, 18, 20
>
> $$n = 7 \Rightarrow \text{Median} = X_{(4)} = 12$$
>
> Because there are **7 data points (odd number)**, the median is located at the **(n + 1) / 2 = 4th position** when the data are arranged in ascending order. Hence, the **4th value**, which is **12**, becomes the **median** — the central value that divides the dataset into two equal parts:
>
> - Lower half: 5, 7, 8
> - Upper half: 15, 18, 20

### 4.1.3 Mode

The **mode** is the **most frequently occurring value** in a dataset. It can be used for **nominal**, **ordinal**, **interval**, or **ratio** data [44].

> **Example**
>
> Data: 3, 4, 4, 5, 6, 6, 6, 7
>
> $$Mode = 6 \text{(because it appears most often)}$$
>
> The most common value in the dataset is **6**.

## 4.2 Appropriate Measure

When analyzing data, selecting the correct measure of central tendency is crucial. The appropriate measure (mean, median, or mode) depends on the **type of data** whether it is categorical, ordered, or numeric. Using the right measure ensures that your analysis accurately reflects the nature and distribution of the data [45].

| Type of Data | Suitable Measure | Explanation |
|---|---|---|
| Nominal | Mode | Data in categories (e.g., color, gender) |
| Ordinal | Median or Mode | Ordered data without equal spacing (e.g., rank, satisfaction level) |
| Interval / Ratio | Mean | Numeric data with meaningful intervals (e.g., income, weight) |

> **Note:**
>
> If the dataset contains **extreme outliers**, use the **median** since it is less affected by extreme values compared to the mean.

## 4.3 Conditional Rule

The choice of which measure of central tendency to use also depends on the **condition or pattern of the data**. Different data shapes and distributions can influence which statistic best represents the center of the dataset [46].

| Data Condition | Recommended Measure |
|---|---|
| Data without outliers and **symmetrical** | **Mean** |

| Data Condition | Recommended Measure |
|---|---|
| Data with **outliers** or **skewed** distribution | **Median** |
| **Categorical** data | **Mode** |
| **Multimodal** data (more than one peak) | **Mode** (can be multi-mode) |

⚠ Explanation:

- When the data is **symmetrical and clean (no outliers)**, the **mean** gives a good overall representation.
- If the data contains **extreme values** or is **skewed**, the **median** is more reliable because it is not affected by those extremes.
- For **categorical variables**, the **mode** identifies the most frequent category.
- In some datasets with multiple peaks, there can be **more than one mode**, indicating several dominant values or groups.

## 4.4 Visualization for CT

Understanding measures of central tendency—**mean**, **median**, and **mode**—is more intuitive when supported by visualizations. Graphical representations such as **histograms** and **boxplots** help reveal the underlying shape, spread, and balance of a dataset. Through these visual tools, we can identify whether the data are **symmetrical**, **skewed**, **categorical**, or **multimodal** [47].

Each visualization provides unique insights:
- **Histograms** show the frequency distribution and how central measures align with data concentration.
- **Boxplots** highlight the median, quartiles, and presence of outliers in a concise format.

In the following subsections, we will explore how central tendency behaves under different conditions using both histogram and boxplot visualizations:

- **Symmetrical and No Outliers** – when data are evenly distributed around the center.
- **Extreme Values (Skewed)** – when outliers pull the mean in one direction.
- **Categorical Variables** – when data represent distinct groups or classes.
- **More Than One Mode** – when data have multiple peaks or centers of concentration.

### 4.4.1 Symmetrical and No outliers

A symmetrical distribution occurs when data values are evenly spread around the center, creating a balanced and bell-shaped pattern. In this case, the **mean**, **median**, and **mode** all fall at or near

the same central point. This indicates that there are no significant **outliers** or **skewness** pulling the data to one side.

In the Figure 4.2, the smooth **density curve** highlights the normal distribution of values, while the vertical lines represent the positions of the mean, median, and mode — all nearly overlapping at the center. Such a distribution is typical for naturally occurring phenomena like height, weight, or measurement errors.

```
library(ggplot2)
# --- Symmetrical data: Perfect bell-shaped (Normal Distribution, no outliers) ---
set.seed(123)
data_sym <- data.frame(value = rnorm(50000, mean = 50, sd = 10))
# --- Compute Mean, Median, Mode ---
mean_val <- mean(data_sym$value)
median_val <- median(data_sym$value)
mode_val <- as.numeric(names(sort(table(round(data_sym$value, 0)),
                                  decreasing = TRUE)[1]))
# --- Visualization (Histogram + Density) ---
ggplot(data_sym, aes(x = value)) +
  geom_histogram(aes(y = after_stat(density)),
                 binwidth = 2,
                 fill = "#5ab4ac",
                 color = "white",
                 alpha = 0.8) +
  geom_density(color = "#2b8cbe", linewidth = 1.3, alpha = 0.9) +
  geom_vline(aes(xintercept = mean_val, color = "Mean"), linewidth = 1.2) +
  geom_vline(aes(xintercept = median_val, color = "Median"),
             linewidth = 1.2, linetype = "dashed") +
  geom_vline(aes(xintercept = mode_val, color = "Mode"),
             linewidth = 1.2, linetype = "dotdash") +
  labs(
    title = "Symmetrical Distribution (No Outliers)",
    subtitle = "Mean, Median, and Mode coincide at the center of the bell curve",
    x = "Value",
    y = "Density",
    color = "Measure"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "bottom"
  )
```

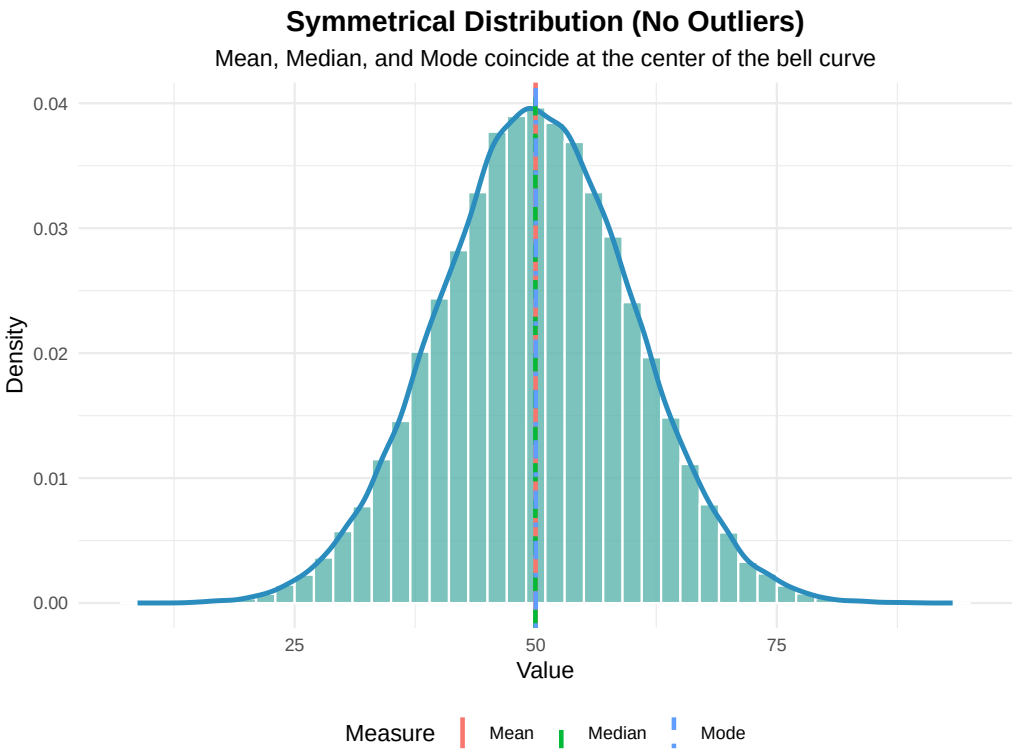



Figure 4.2: Symmetrical Distribution (No Outliers)

Explanation:

A **symmetrical distribution** represents a balanced dataset where values are evenly distributed around the central point. This pattern forms the classic **bell-shaped curve**, also known as a **normal distribution**. In such cases, the **mean**, **median**, and **mode** are equal or nearly identical, reflecting perfect equilibrium in the data.

**Key Interpretations**

- **Balance Around the Center:** Data are distributed evenly on both sides of the center, showing no bias toward higher or lower values.
- **Equality of Central Measures:** The mean, median, and mode overlap or align closely, indicating that the dataset is centered without distortion from extreme values.
- **Absence of Skewness and Outliers:* There are no outliers pulling the data to one side, and the distribution is neither left- nor right-skewed. This results in a stable and predictable shape.
- **Predictable Shape — Bell Curve:** The histogram and smooth density curve form a bell shape, where most values cluster near the center, and frequencies gradually taper off toward both tails.

**Statistical Implication**

Such a symmetrical pattern satisfies many **classical statistical assumptions**, making it foundational for various **parametric analyses** such as:

- **t-tests**
- **ANOVA**
- **Linear regression**

Because the data follow a normal distribution, inferential analyses become **more valid, reliable, and stable**, as deviations and sampling errors are minimized.

**Real-World Examples**

Symmetrical, bell-shaped distributions commonly appear in:

- **Human characteristics** (e.g., height, weight, IQ)
- **Natural phenomena** (e.g., measurement errors, biological variation)
- **Academic performance** (e.g., exam scores from large populations)

### 4.4.2 Extreme Values (Skewed)

A skewed distribution occurs when data values are not symmetrically distributed around the center — meaning one tail of the distribution is longer or more stretched than the other. This skewness is often caused by extreme values (outliers) that pull the mean toward one direction, while the median and mode remain closer to the peak of the data.

When a dataset contains extreme high or low values, the distribution becomes positively skewed (right-skewed) or negatively skewed (left-skewed). These distortions affect the position of central tendency measures and provide valuable insight into the underlying data behavior. In the Figure 4.3 plot below, we can observe how a few extreme values shift the mean away from the main cluster of data, creating an asymmetrical shape.

```
library(ggplot2)
# --- Right-skewed data (with extreme values) ---
set.seed(123)
data_skew <- data.frame(value = c(rgamma(4800, shape = 2, scale = 10),
                                  rnorm(200, mean = 200, sd = 5)))
# --- Compute Mean, Median, Mode ---
mean_val <- mean(data_skew$value)
median_val <- median(data_skew$value)
mode_val <- as.numeric(names(sort(table(round(data_skew$value, 0)),
                                  decreasing = TRUE)[1]))
# --- Visualization (Histogram + Density) ---
ggplot(data_skew, aes(x = value)) +
geom_histogram(aes(y = after_stat(density)),
binwidth = 5,
fill = "#5ab4ac",
color = "white",
```

```
alpha = 0.8) +
geom_density(color = "#2b8cbe", linewidth = 1.3, alpha = 0.9) +
geom_vline(aes(xintercept = mean_val, color = "Mean"), linewidth = 1.2) +
geom_vline(aes(xintercept = median_val, color = "Median"), linewidth = 1.2,
           linetype = "dashed") +
geom_vline(aes(xintercept = mode_val, color = "Mode"), linewidth = 1.2,
           linetype = "dotdash") +
labs(
title = "Right-Skewed Distribution (With Extreme Values)",
subtitle = "Mean is pulled toward the extreme high values due to skewness",
x = "Value",
y = "Density",
color = "Measure"
) +
theme_minimal(base_size = 13) +
theme(
plot.title = element_text(face = "bold", hjust = 0.5),
plot.subtitle = element_text(hjust = 0.5),
legend.position = "bottom"
)
```

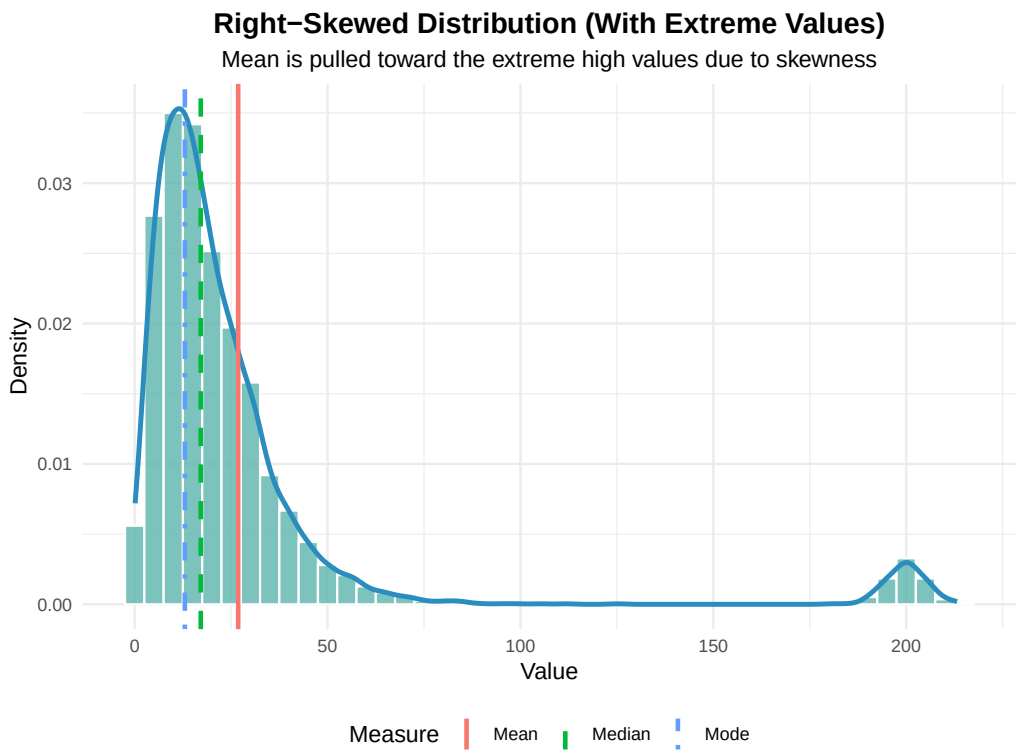



Figure 4.3: Right-Skewed Distribution (With Extreme Values)

⚠ Explanation:

A skewed distribution indicates that the dataset is asymmetrical, meaning the data do not fall evenly around the central point. The presence of extreme values (outliers) causes this imbalance, pulling one side of the distribution's tail farther than the other.

## Types of Skewness

- **Positively Skewed (Right-Skewed):** The tail extends to the right, showing that a small number of large values are stretching the mean upward (Order: Mode < Median < Mean).
- **Negatively Skewed (Left-Skewed):** The tail extends to the left, indicating that very small values pull the mean downward (Order: Mean < Median < Mode).

### Key Interpretations

- **Effect of Extreme Values:** Outliers on one end distort the balance of the distribution, shifting the mean away from the center.
- **Separation of Central Measures:** The mean, median, and mode no longer align; their spacing reveals the degree of skewness.
- **Asymmetrical Shape:** The histogram shows one side tapering more gradually, confirming the directional bias in the data.
- **Impact on Statistical Assumptions:** Skewed data violate normality assumptions required in many parametric tests.

### Statistical Implication

Skewness can influence data interpretation and analysis validity, especially when using parametric methods such as t-tests, ANOVA, or linear regression, which assume normality. In such cases, analysts often:

- Apply data transformation (e.g., log, square root)
- Use non-parametric tests (e.g., Mann–Whitney U, Kruskal–Wallis)
- Identify and handle outliers explicitly

### Real-World Examples

Right- or left-skewed distributions commonly appear in:

- **Income and wealth data:** A few individuals earn far more than most (right-skewed).
- **Time-to-failure or lifespan data:** Many items fail early, with fewer lasting very long (left-skewed).
- **Sales and transaction data:** A small number of customers may account for extremely high purchase amounts.
- **Environmental data:** Some readings, like pollution concentration, exhibit right-skew due to rare spikes.

### 4.4.3 Categorical Variables

Categorical variables divide data into distinct groups or categories. When combined with a numerical variable, we can analyze how the distribution of numerical values differs across categories. A boxplot is an excellent visualization for this purpose — it shows the median, quartiles, range, and outliers within each group.

For example, in the chart below (Figure 4.4), each box represents a product category, while the vertical axis shows the distribution of sales values within that category.

```
# ============================================================
# Categorical Variables - Boxplot Visualization (Fixed & Varied Distributions)
# ============================================================
library(ggplot2)
library(dplyr)

set.seed(123)

# --- Create category structure ---
categories <- c("Electronics", "Clothing", "Home", "Beauty", "Sports")
sales_data <- data.frame(
  ProductCategory = sample(
    categories, 500, replace = TRUE,
    prob = c(0.25, 0.30, 0.20, 0.15, 0.10)
  )
)

# --- Generate different distributions per category correctly ---
sales_data <- sales_data %>%
  group_by(ProductCategory) %>%
  mutate(
    TotalSales = case_when(
      ProductCategory == "Electronics" ~ rnorm(n(), mean = 120, sd = 20),          # normal, s
      ProductCategory == "Clothing"    ~ rlnorm(n(), meanlog = 4.5, sdlog = 0.4), # right-ske
      ProductCategory == "Home"        ~ runif(n(), min = 60, max = 150),         # uniform
      ProductCategory == "Beauty"      ~ rexp(n(), rate = 1/70),                  # exponenti
      ProductCategory == "Sports"      ~ rnorm(n(), mean = 90, sd = 35)           # wide spre
    )
  ) %>%
  ungroup()

# --- Visualization: Boxplot by Category ---
ggplot(sales_data, aes(x = ProductCategory, y = TotalSales, fill = ProductCategory)) +
  geom_boxplot(
    alpha = 0.8, color = "gray30",
    outlier.colour = "red", outlier.shape = 16, outlier.size = 2
  ) +
  labs(
    title = "Boxplot of Sales Distribution by Product Category",
    subtitle = "Each category displays a unique distribution pattern of total sales",
```

```
    x = "Product Category",
    y = "Total Sales (in units)",
    fill = "Category"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "none"
  )
```

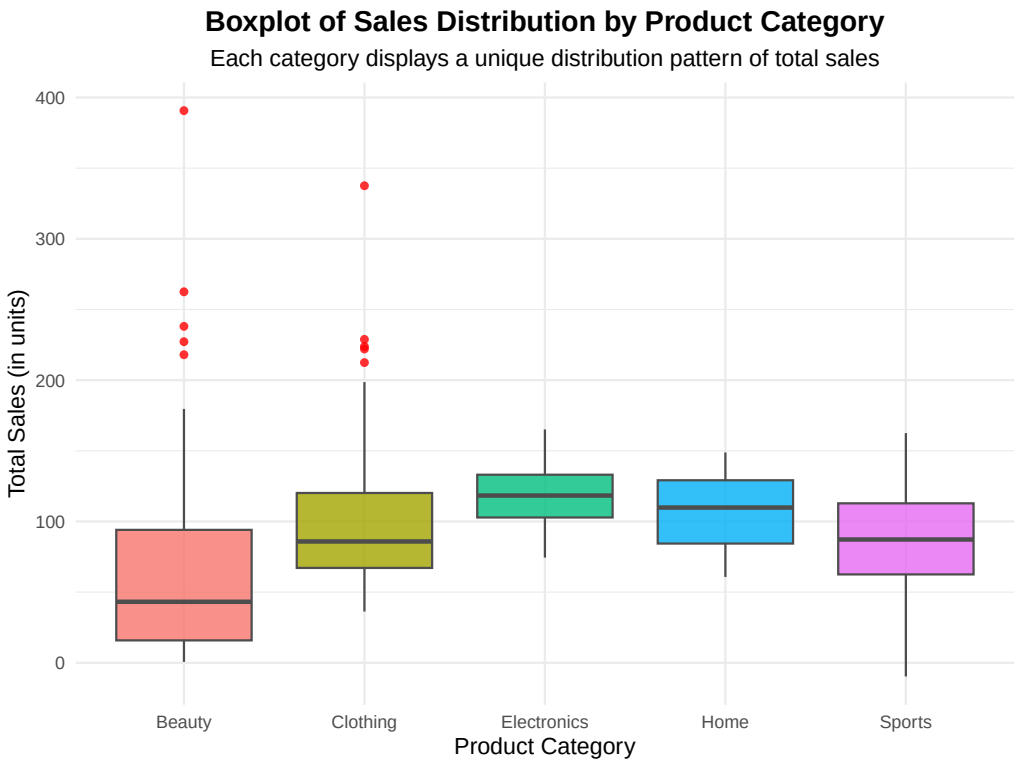
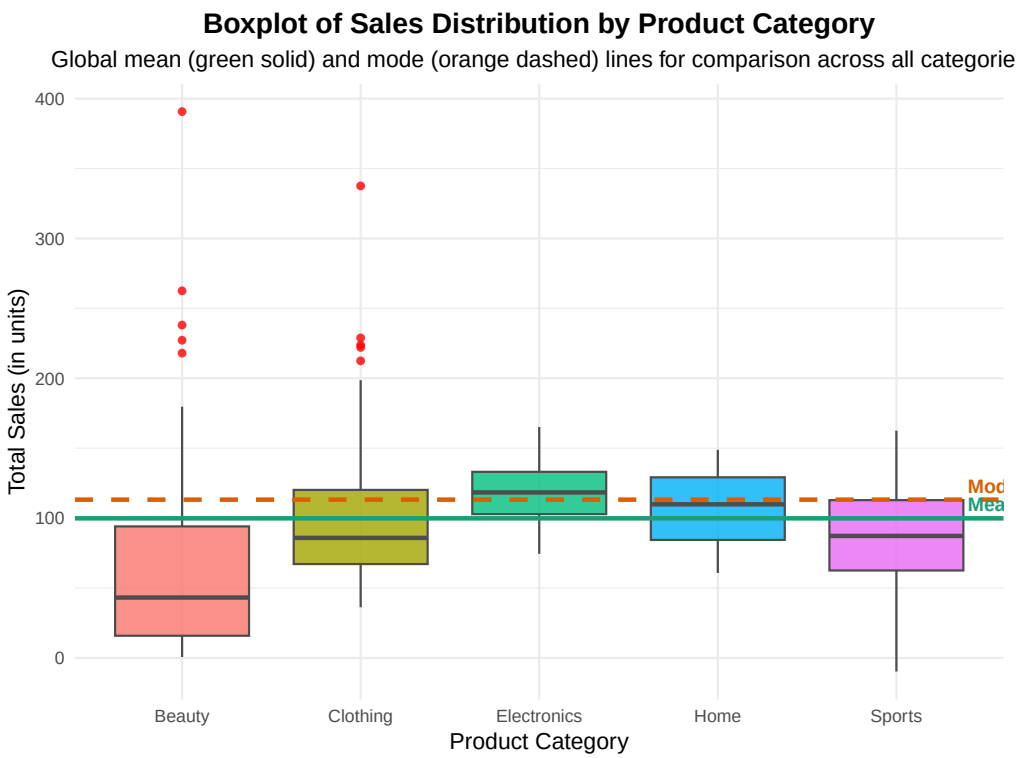



Figure 4.4: Boxplot of Sales Distribution by Product Category

```
# ============================================================
# Boxplot of Sales by Product Category - with Global Mean & Mode Lines
# ============================================================
library(ggplot2)
library(dplyr)

set.seed(123)

# --- Create category structure ---
categories <- c("Electronics", "Clothing", "Home", "Beauty", "Sports")
sales_data <- data.frame(
```

```
  ProductCategory = sample(
    categories, 500, replace = TRUE,
    prob = c(0.25, 0.30, 0.20, 0.15, 0.10)
  )
)

# --- Generate different distributions per category ---
sales_data <- sales_data %>%
  group_by(ProductCategory) %>%
  mutate(
    TotalSales = case_when(
      ProductCategory == "Electronics" ~ rnorm(n(), mean = 120, sd = 20),
      ProductCategory == "Clothing"    ~ rlnorm(n(), meanlog = 4.5, sdlog = 0.4),
      ProductCategory == "Home"        ~ runif(n(), min = 60, max = 150),
      ProductCategory == "Beauty"      ~ rexp(n(), rate = 1/70),
      ProductCategory == "Sports"      ~ rnorm(n(), mean = 90, sd = 35)
    )
  ) %>%
  ungroup()

# --- Compute global mean and mode ---
mean_val <- mean(sales_data$TotalSales, na.rm = TRUE)

# Estimate mode using kernel density (works for continuous data)
dens <- density(sales_data$TotalSales, na.rm = TRUE)
mode_val <- dens$x[which.max(dens$y)]

# --- Visualization ---
ggplot(sales_data, aes(x = ProductCategory, y = TotalSales, fill = ProductCategory)) +
  geom_boxplot(
    alpha = 0.8, color = "gray30",
    outlier.colour = "red", outlier.shape = 16, outlier.size = 2
  ) +
  # Add mean line
  geom_hline(
    yintercept = mean_val, color = "#1b9e77", linewidth = 1.2
  ) +
  # Add mode line
  geom_hline(
    yintercept = mode_val, color = "#d95f02", linewidth = 1.2, linetype = "dashed"
  ) +
  # Annotate lines
  annotate(
    "text", x = 5.4, y = mean_val, label = sprintf("Mean = %.1f", mean_val),
    color = "#1b9e77", hjust = 0, vjust = -0.5, fontface = "bold", size = 3.8
  ) +
  annotate(
    "text", x = 5.4, y = mode_val, label = sprintf("Mode = %.1f", mode_val),
    color = "#d95f02", hjust = 0, vjust = -0.5, fontface = "bold", size = 3.8
```

```
) +
labs(
  title = "Boxplot of Sales Distribution by Product Category",
  subtitle = "Global mean (green solid) and mode (orange dashed) lines for comparison ac
  x = "Product Category",
  y = "Total Sales (in units)"
) +
theme_minimal(base_size = 13) +
theme(
  plot.title = element_text(face = "bold", hjust = 0.5),
  plot.subtitle = element_text(hjust = 0.5),
  legend.position = "none"
)
```




**Explanation:**

A categorical variable divides data into distinct groups or categories — for example, product types, departments, or regions — while a numerical variable measures quantitative outcomes such as sales, profit, or ratings. When visualized using a boxplot, the relationship between these two types of variables becomes clear, showing how the numerical data are distributed within each category.

**Key Interpretations**

- **Median (Center Line):** Represents the central value of sales within each category, showing which category tends to sell more or less.
- **Interquartile Range (IQR):** The height of each box shows the middle 50% of the data — wider boxes indicate greater variability in sales.
- **Whiskers and Outliers:** The vertical lines (whiskers) represent typical sales ranges, while the red dots highlight outliers (unusually high or low values).
- **Comparison Across Categories:** Different box heights and positions indicate variation in both central tendency and spread among product categories.

**Statistical Implications**

Boxplots of categorical variables are valuable for:

- Detecting differences in distribution among groups.
- Identifying skewness and outliers within each category.
- Assessing variability and central tendency visually without relying on complex statistical summaries.

**Real-World Applications**

This approach is essential in:

- **Business analytics:** Comparing sales or profits across product lines.
- **Healthcare:** Comparing recovery times or satisfaction scores across hospitals.
- **Education:** Comparing test scores across schools or departments.

By visualizing categorical variables with boxplots, analysts can quickly detect differences between groups, guide deeper statistical testing, and support data-driven decisions.

### 4.4.4 More Than One Mode

In many real-world datasets, the distribution of values does not always form a single, smooth peak. Instead, some datasets exhibit two or more distinct peaks, known as multiple modes. Each mode represents a cluster where values tend to concentrate — meaning that the data have several regions of high frequency rather than one central location.

In the following histogram (see, Figure 4.5), we will observe a bimodal distribution where two separate peaks appear clearly. This illustrates how the histogram can reveal hidden structure in the data that simple summary statistics, like the mean or median, might overlook.

```
# ============================================================
# More Than One Mode - Bimodal Distribution Visualization
# ============================================================
library(ggplot2)
library(dplyr)
set.seed(123)
```

```r
# --- Generate Bimodal Data (two peaks) ---
# Combine two normal distributions with different means
data_bimodal <- data.frame(
  value = c(
    rnorm(2500, mean = 40, sd = 6),   # first cluster
    rnorm(2500, mean = 70, sd = 6)    # second cluster
  )
)
# --- Compute Summary Statistics ---
mean_val   <- mean(data_bimodal$value)
median_val <- median(data_bimodal$value)
# --- Visualization: Histogram + Density Curve ---
ggplot(data_bimodal, aes(x = value)) +
  geom_histogram(
    aes(y = after_stat(density)),
    bins = 40, fill = "#74a9cf",
    color = "white", alpha = 0.8
  ) +
  geom_density(color = "#0570b0", linewidth = 1.3, alpha = 0.9) +
  geom_vline(aes(xintercept = mean_val, color = "Mean"), linewidth = 1.2) +
  geom_vline(aes(xintercept = median_val, color = "Median"),
             linewidth = 1.2, linetype = "dashed") +
  labs(
    title = "Bimodal Distribution (More Than One Mode)",
    subtitle = "Two distinct peaks represent different groups or subpopulations",
    x = "Value",
    y = "Density",
    color = "Measure"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "bottom"
  )
```

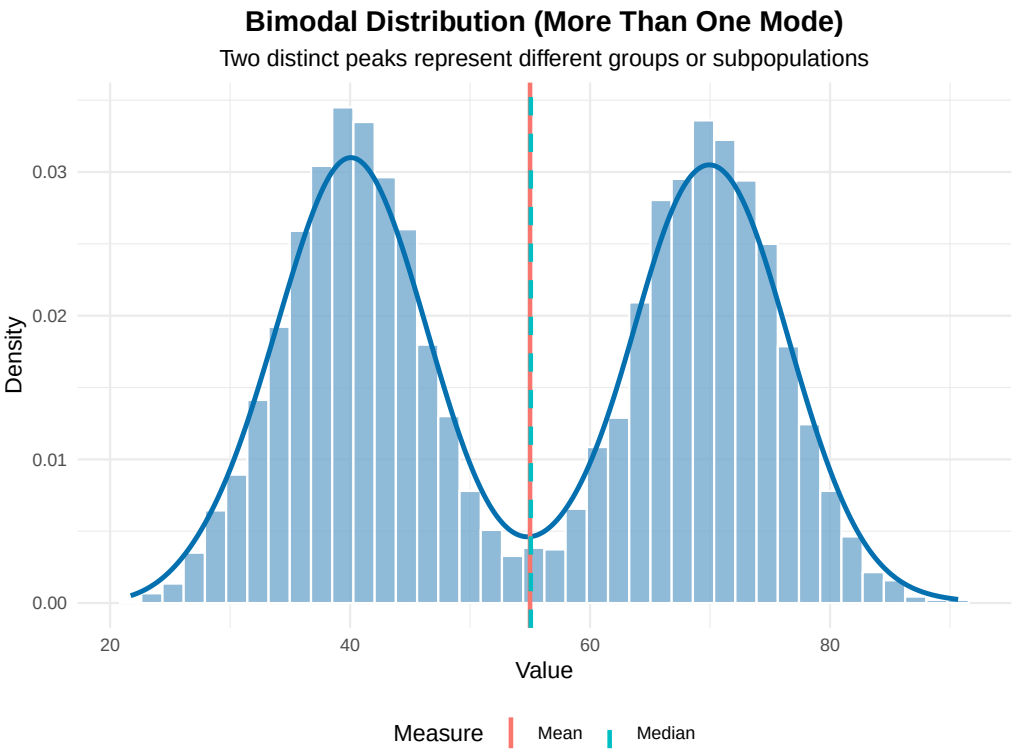



Figure 4.5: Bimodal Distribution (More Than One Mode)

Unlike histograms, boxplots do not display the exact number of peaks, but they clearly show that the data are not symmetrically distributed — for example, the median line may be off-center, and the whiskers might extend unevenly to one side. Together, the histogram and boxplot provide complementary insights:

- the histogram reveals the overall shape (and multiple modes),
- while the boxplot emphasizes the spread and skewness of the data.

In the following visualization (see, Figure 4.6), the boxplot helps us interpret how a bimodal dataset behaves in terms of variation, central value, and outliers, reinforcing the insights gained from the histogram.

```
# ============================================================
# Boxplot Representation - Bimodal Distribution
# ============================================================
library(ggplot2)
library(dplyr)
set.seed(123)
# --- Generate Bimodal Data (same as histogram) ---
data_bimodal <- data.frame(
value = c(
rnorm(2500, mean = 40, sd = 6),   # first cluster
```

```
rnorm(2500, mean = 70, sd = 6)    # second cluster
)
)
# --- Compute Summary Statistics ---
mean_val   <- mean(data_bimodal$value)
median_val <- median(data_bimodal$value)
# --- Visualization: Boxplot ---
ggplot(data_bimodal, aes(x = "", y = value)) +
geom_boxplot(
fill = "#74a9cf",
color = "gray30",
outlier.colour = "#fb6a4a",
outlier.shape = 16,
outlier.size = 2,
width = 0.3
) +
geom_hline(aes(yintercept = mean_val, color = "Mean"), linewidth = 1.2) +
geom_hline(aes(yintercept = median_val, color = "Median"), linewidth = 1.2,
           linetype = "dashed") +
labs(
title = "Boxplot of Bimodal Distribution (More Than One Mode)",
subtitle = "Wider spread indicates data concentration around two regions",
x = NULL,
y = "Value",
color = "Measure"
) +
scale_color_manual(values = c("Mean" = "#0570b0", "Median" = "#ff7f00")) +
theme_minimal(base_size = 13) +
theme(
plot.title = element_text(face = "bold", hjust = 0.5),
plot.subtitle = element_text(hjust = 0.5),
axis.text.x = element_blank(),
legend.position = "bottom"
)
```

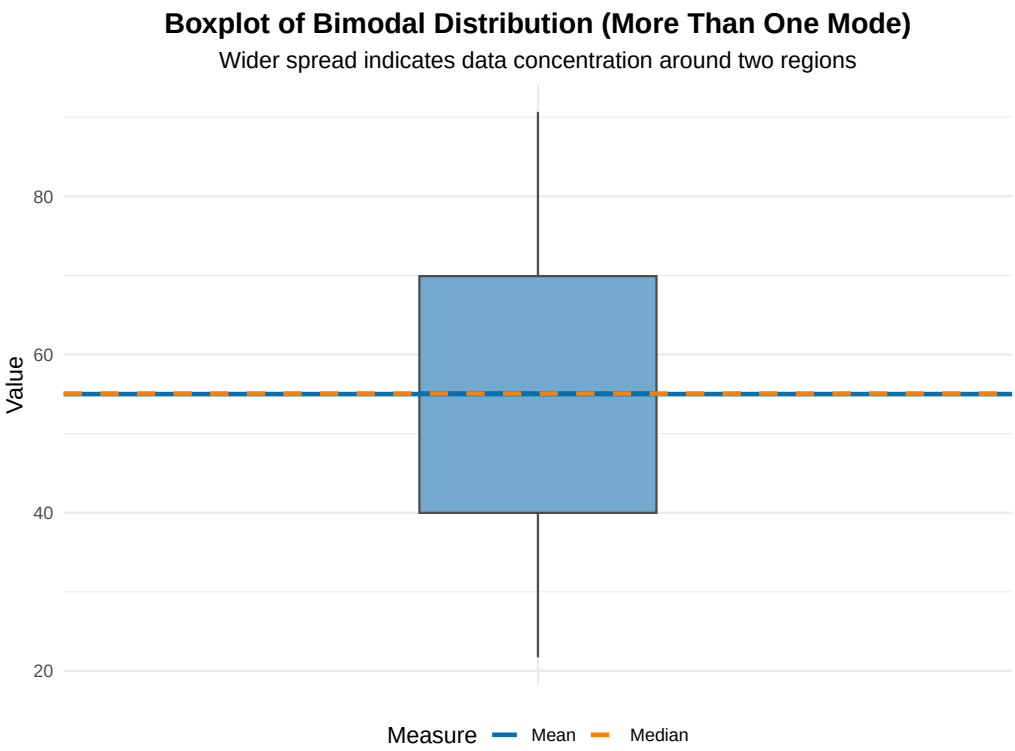



Figure 4.6: Bimodal Distribution (More Than One Mode)

> **Explanation:**
>
> A boxplot cannot explicitly display two peaks (bimodal pattern), because:
>
> - The boxplot only summarizes data statistically (using the five-number summary: minimum, first quartile, median, third quartile, and maximum).
> - It does not represent the shape of the distribution (e.g., how many peaks or modes exist).

```r
# ============================================================
# Enhanced Violin + Boxplot - Bimodal Distribution
# ============================================================
library(ggplot2)
library(ggtext)
set.seed(123)

data_bimodal <- data.frame(
  value = c(
    rnorm(2500, mean = 40, sd = 6),
    rnorm(2500, mean = 70, sd = 6)
  )
)
```

```
# --- Create the plot ---
ggplot(data_bimodal, aes(x = "", y = value)) +
  # Gradient violin to show smooth density
  geom_violin(
    aes(fill = stat(y)),
    color = "gray30",
    alpha = 0.8,
    width = 1.1,
    linewidth = 0.6
  ) +
  scale_fill_gradient(
    low = "#c6dbef", high = "#08306b", name = "Density"
  ) +
  # Overlay boxplot
  geom_boxplot(
    width = 0.12,
    fill = "#fdd0a2",
    color = "gray25",
    outlier.colour = "#fb6a4a",
    outlier.shape = 16,
    outlier.size = 2
  ) +
  # Annotate the two modes
  annotate(
    "text", x = 1.15, y = 40, label = "First Mode $\\approx$ 40",
    color = "#045a8d", size = 4, fontface = "bold", hjust = 0
  ) +
  annotate(
    "text", x = 1.15, y = 70, label = "Second Mode $\\approx$ 70",
    color = "#d94801", size = 4, fontface = "bold", hjust = 0
  ) +
  annotate(
    "curve",
    x = 1.05, xend = 1.0, y = 42, yend = 45,
    curvature = 0.3, color = "#045a8d", arrow = arrow(length = unit(0.15, "cm"))
  ) +
  annotate(
    "curve",
    x = 1.05, xend = 1.0, y = 68, yend = 65,
    curvature = -0.3, color = "#d94801", arrow = arrow(length = unit(0.15, "cm"))
  ) +
  labs(
    title = "Violin + Boxplot of Bimodal Distribution",
    subtitle = "The violin shape reveals two clear concentration regions around
    <b style='color:#045a8d;'>40</b> and <b style='color:#d94801;'>70</b>",
    x = NULL,
    y = "Value",
    fill = "Density"
  ) +
```

```
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    plot.subtitle = element_markdown(hjust = 0.5, size = 12),
    axis.text.x = element_blank(),
    legend.position = "none",
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank(),
    plot.background = element_rect(fill = "#f9fbfd", color = NA)
  )
```

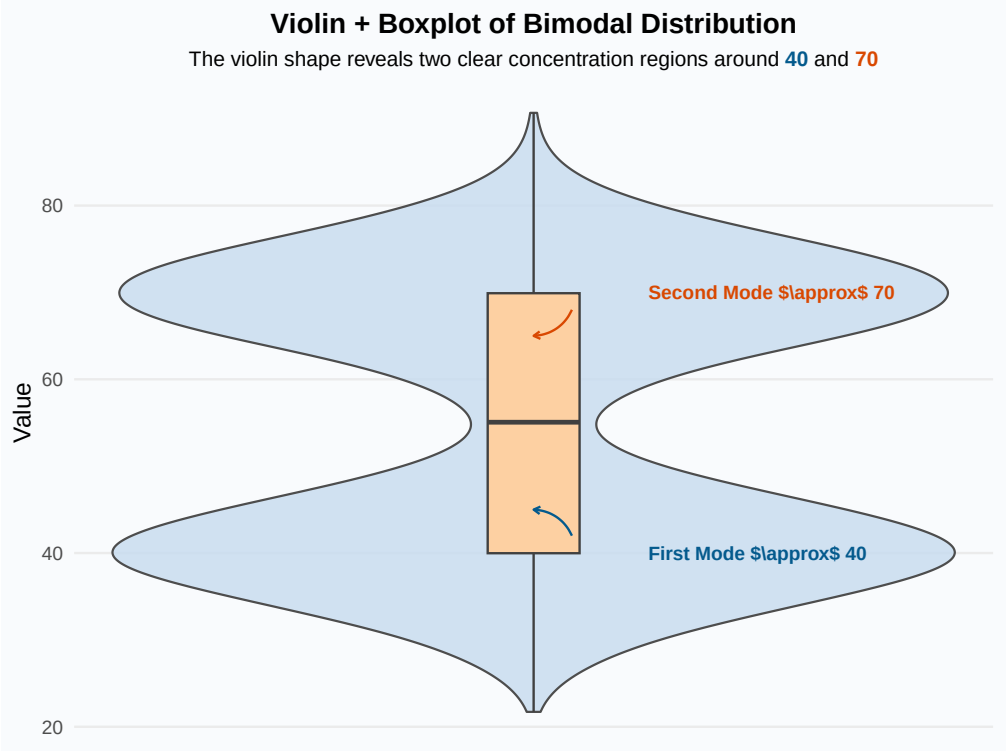



Figure 4.7: Violin + Boxplot of Bimodal Distribution

Explanation:

A bimodal distribution occurs when a dataset has two distinct peaks (modes), meaning there are two dominant groups of values around different centers. Unlike a normal distribution that has one central peak, a bimodal shape suggests that the data may come from two different populations or underlying processes combined into one dataset.

**Key Interpretations**

- **Two Peaks (Modes):** Each peak represents a cluster of frequently occurring values — often caused by two subgroups with different characteristics.
- **Mean and Median:** These measures may fall between the two modes, failing to represent either group accurately.
- **Spread and Overlap:** The distance between peaks and the overlap between them indicate how distinct or similar the two groups are.
- **Potential Mixture of Populations:** Bimodality is a strong clue that the dataset may not be homogeneous.

**Statistical Implications**

- Classical measures like mean and standard deviation can be misleading, since they ignore multimodal structure.
- Analysts should consider segmenting the data (e.g., clustering or grouping) before running inferential tests.
- Identifying multiple modes often leads to insightful segmentation — discovering hidden subgroups within the data.

## 4.5 Dataset

Copy CSV Search:

Interactive Table: Customer Purchase Data

| CustomerID | Age | Gender | StoreLocation | ProductCategory | TotalPurchase | NumberOfVisits | FeedbackScore |
|---|---|---|---|---|---|---|---|
| 1 | 32 | M | West | Electronics | 528 | 4 | 1 |
| 2 | 37 | F | South | Books | 72 | 4 | 5 |
| 3 | 63 | M | West | Electronics | 327 | 4 | 2 |
| 4 | 41 | M | North | Sports | 391 | 7 | 1 |
| 5 | 42 | F | East | Electronics | 514 | 7 | 5 |
| 6 | 66 | F | East | Sports | 381 | 6 | 3 |
| 7 | 47 | M | East | Sports | 510 | 5 | 1 |
| 8 | 21 | F | South | Clothing | 102 | 4 | 2 |
| 9 | 30 | F | North | Sports | 559 | 2 | 2 |
| 10 | 33 | M | South | Books | 27 | 5 | 2 |

Showing 1 to 10 of 200 entries Previous 1 2 3 4 5 … 20 Next

# References

# Chapter 5

# Statistical Dispersion

While Central Tendency identifies the "middle" of a dataset, **Measures of Dispersion/Variability** describe how widely the values are spread around that center. In other words, dispersion quantifies the degree of variability or diversity within the data. Two datasets can share the same average, yet their distributions may look completely different—one tightly clustered, the other broadly scattered.

**Watch here:** Statistical Dispersion

By combining Central Tendency with these measures of dispersion (see, Figure 5.1), readers gain both numerical and visual insights, enabling a more accurate and holistic interpretation of their data [48]–[51].

## 5.1 Range

The *range* is the simplest measure of dispersion, representing the difference between the largest and smallest observations in a dataset. It provides a quick sense of how spread out the data are [52].

**Formula:**

$$Range = X_{max} - X_{min}$$

A larger range indicates greater variability among the data values, while a smaller range suggests that the data are more concentrated around the mean. The range is easy to compute and understand, making it a useful measure for providing a quick and rough estimate of how widely the data are spread. However, it has notable limitations: it is highly sensitive to outliers and does not take into account the distribution of values between the smallest and largest observations [53], [54].

**Example:**

A researcher measures the **systolic blood pressure reduction (in mmHg)** of five patients after taking **Drug A**:

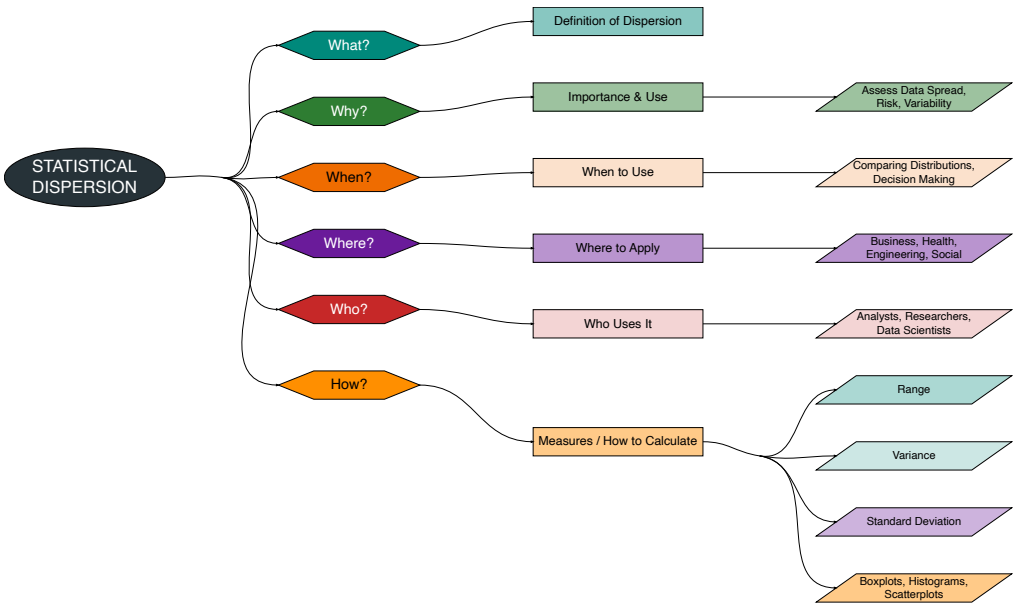



Figure 5.1: Statistical Dispersion 5W+1H

$$[45, 52, 49, 47, 55]$$

We can compute the *range* of the blood pressure reduction values, as the following:

$$Range = X_{max} - X_{min} = 55 - 45 = 10$$

The reductions vary by **10 mmHg**, indicating relatively low variability among patients.

## 5.2 Variance

*Variance* measures the average of the squared deviations from the mean. It quantifies how much each data point differs from the mean, capturing the degree of spread in the dataset.

**Formulas:**

- For a population:
$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$
- For a sample:
$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

A higher variance indicates that the data points are more widely spread from the mean, while a lower variance suggests that they are clustered more closely together. Variance takes into account all data points in a dataset, not just the extremes, and serves as the foundation for more advanced statistical measures such as standard deviation and ANOVA. However, because variance is expressed in squared units, it can be less intuitive to interpret directly. Additionally, it is sensitive to extreme values, which can disproportionately affect the measure of variability [53], [54].

**Example:**

We can compute the **sample variance** of the blood pressure reduction values, as the following:

- **Step One:** Compute the mean,

$$\bar{X} = \frac{46 + 50 + 54 + 48 + 52}{5} = \frac{250}{5} = 50$$

- **Step Two:** Compute each squared deviation,

| $X_i$ | $(X_i - \bar{X})$ | $(X_i - \bar{X})^2$ |
|---|---|---|
| 46 | -4 | 16 |
| 50 | 0 | 0 |

| $X_i$ | $(X_i - \bar{X})$ | $(X_i - \bar{X})^2$ |
|---|---|---|
| 54 | 4 | 16 |
| 48 | -2 | 4 |
| 52 | 2 | 4 |

- **Step Three:** Compute the sample variance:

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

$$s^2 = \frac{16 + 0 + 16 + 4 + 4}{5 - 1} = \frac{40}{4} = 10$$

The **sample variance** is **10 (mmHg²)**, meaning that, on average, the squared deviations of blood pressure reductions from the mean are 10 units.

## 5.3 Standard Deviation

The *standard deviation* (SD) is the square root of the variance. It measures the average distance of each data point from the mean and is expressed in the same units as the original data.

**Formulas:**

- For a population:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

- For a sample:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

A low standard deviation indicates that the data points are close to the mean, reflecting low variability within the dataset, while a high standard deviation shows that the data points are more widely dispersed, indicating higher variability. One of the main advantages of standard deviation is that it is expressed in the same units as the original data, making it easier to interpret compared to variance. It is also widely used in both descriptive and inferential statistics for assessing data consistency and reliability. However, standard deviation is influenced by outliers, which can distort the measure of spread, and it assumes that the data distribution is approximately normal for its interpretation to be most meaningful [55].

**Example:**

Recall that the **sample variance** ($s^2$) was:

$$s^2 = 10$$

The **standard deviation** ($s$) is the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{10} \approx 3.16$$

The **sample standard deviation** is **3.16 mmHg**, which means that, on average, each patient's blood pressure reduction differs from the mean by about **3.16 mmHg**.

## 5.4 Study Cases

A clinical study was conducted to evaluate the effectiveness and consistency of three different antihypertensive drugs—Drug A, Drug B, and Drug C—in lowering patients' blood pressure. Each group of patients received one type of drug for four weeks. The goal was to reduce systolic blood pressure (SBP) to around 120 mmHg, which is considered normal according to the World Health Organization (WHO) and the American Heart Association (AHA) guidelines. Although the mean reduction in blood pressure for all three drugs is approximately 50 mmHg, the variability in response differs significantly. This variation reflects how consistent or scattered the treatment effects are among patients. Let consider thi dataset (**tab-dataset-bab51?**)

Copy CSV Search:

| | PatientID | Drug | BP_Reduction |
|---|---|---|---|
| 1 | 1 | Drug A | 46.7455922240579 |
| 2 | 2 | Drug A | 48.3970830940256 |
| 3 | 3 | Drug A | 57.3415120275645 8 |
| 4 | 4 | Drug A | 49.9005124139418 5 |
| 5 | 5 | Drug A | 50.1944091326237 |
| 6 | 6 | Drug A | 58.1232953912353 7 |
| 7 | 7 | Drug A | 51.8525514867649 8 |
| 8 | 8 | Drug A | 43.2226642837862 9 |
| 9 | 9 | Drug A | 46.1137061973513 3 |
| 10 | 10 | Drug A | 47.3196606063191 8 |

Showing 1 to 10 of 300 entries Previous 1 2 3 4 5 … 30 Next

Graphs help visualize the variability in treatment effects among the three drugs:

- Boxplots show the spread and outliers of blood pressure reduction.
- Histograms reveal how reductions are distributed among patients.
- Scatterplots illustrate how responses vary with other factors.

These visuals highlight that Drug A has consistent effects, Drug B shows some outliers, and Drug C displays a wider, skewed variation.

### 5.4.1 Boxplots

Before analyzing the differences among Drug A, Drug B, and Drug C, it is important to understand how boxplots represent data dispersion. A boxplot provides a compact visual summary of a dataset through five key statistics: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum, etc.

```
# Read dataset above then calculate
# Summary statistics to verify means are exactly 50
drug_summary <- drug_data %>%
  group_by(Drug) %>%
  summarise(
```

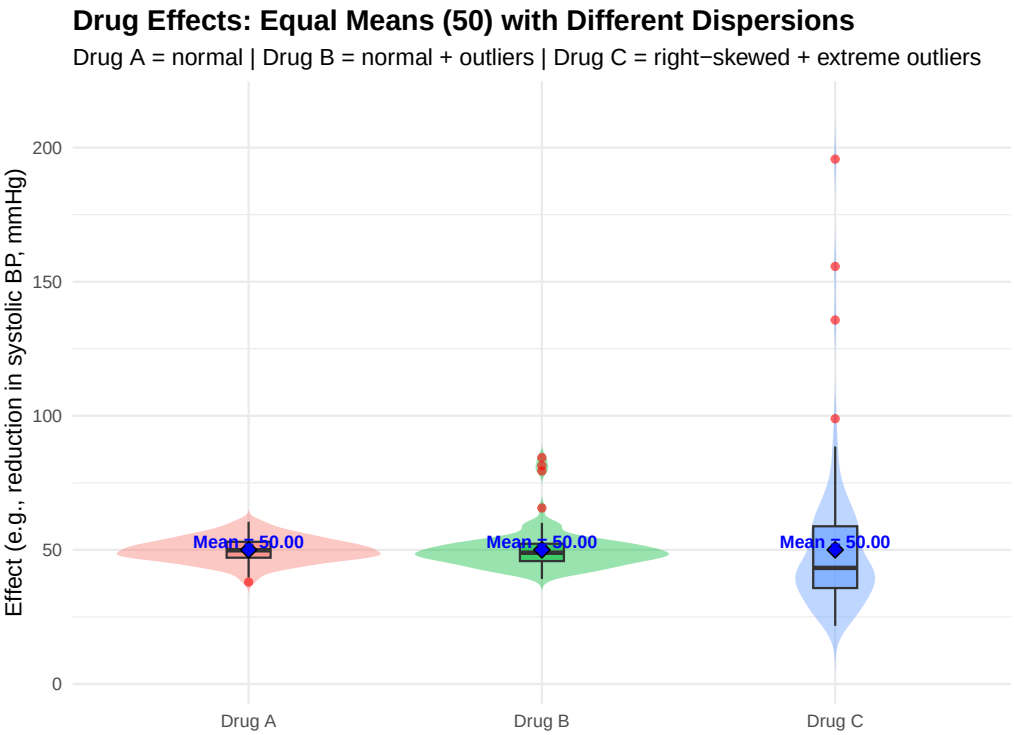
```
    Mean = mean(BP_Reduction),
    Min = min(BP_Reduction),
    Max = max(BP_Reduction),
    Range = Max - Min,
    Variance = var(BP_Reduction),
    SD = sd(BP_Reduction)
  )

# ============================================================
# Display interactive table
# ============================================================
datatable(
  drug_summary %>%
    mutate(across(where(is.numeric), ~round(., 2))),  # numeric to 2 decimals
  options = list(
    dom = 't',         # show only the table
    paging = FALSE,    # disable pagination
    ordering = FALSE  # disable sorting
  ),
  rownames = FALSE
)
```

| Drug | Mean | Min | Max | Range | Variance | SD |
|---|---|---|---|---|---|---|
| Drug A | 50 | 38 | 60.48 | 22.48 | 20.83 | 4.56 |
| Drug B | 50 | 39.15 | 84.42 | 45.27 | 54.29 | 7.37 |
| Drug C | 50 | 21.67 | 195.7 | 174.03 | 657.49 | 25.64 |

The box represents the interquartile range (IQR = Q3 − Q1), showing where the middle 50% of data points lie. The line inside the box marks the median, while the "whiskers" extend to the smallest and largest values within 1.5 × IQR. Any points beyond the whiskers are plotted individually as outliers, indicating unusually high or low observations.

```
# Plot: violin + boxplot with mean annotation --------
library(ggplot2)

ggplot(drug_data, aes(x = Drug, y = BP_Reduction, fill = Drug)) +
  # Violin plot for full distribution
  geom_violin(alpha = 0.4, trim = FALSE, color = NA) +
  # Boxplot overlay (narrower width)
  geom_boxplot(width = 0.15, outlier.color = "red", alpha = 0.6) +
  # Mean point
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "blue", color = "b
  # Mean label
  geom_text(
    data = drug_summary,
    aes(x = Drug, y = Mean + 3,
        label = paste0("Mean = ", formatC(Mean, digits = 2, format = "f"))),
    color = "blue", size = 3.5, fontface = "bold", inherit.aes = FALSE
  ) +
  labs(
    title = "Drug Effects: Equal Means (50) with Different Dispersions",
    subtitle = "Drug A = normal | Drug B = normal + outliers | Drug C = right-skewed + ext
```

```
    x = "",
    y = "Effect (e.g., reduction in systolic BP, mmHg)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    legend.position = "none",
    plot.title = element_text(face = "bold")
  )
```




**Interpretation:**

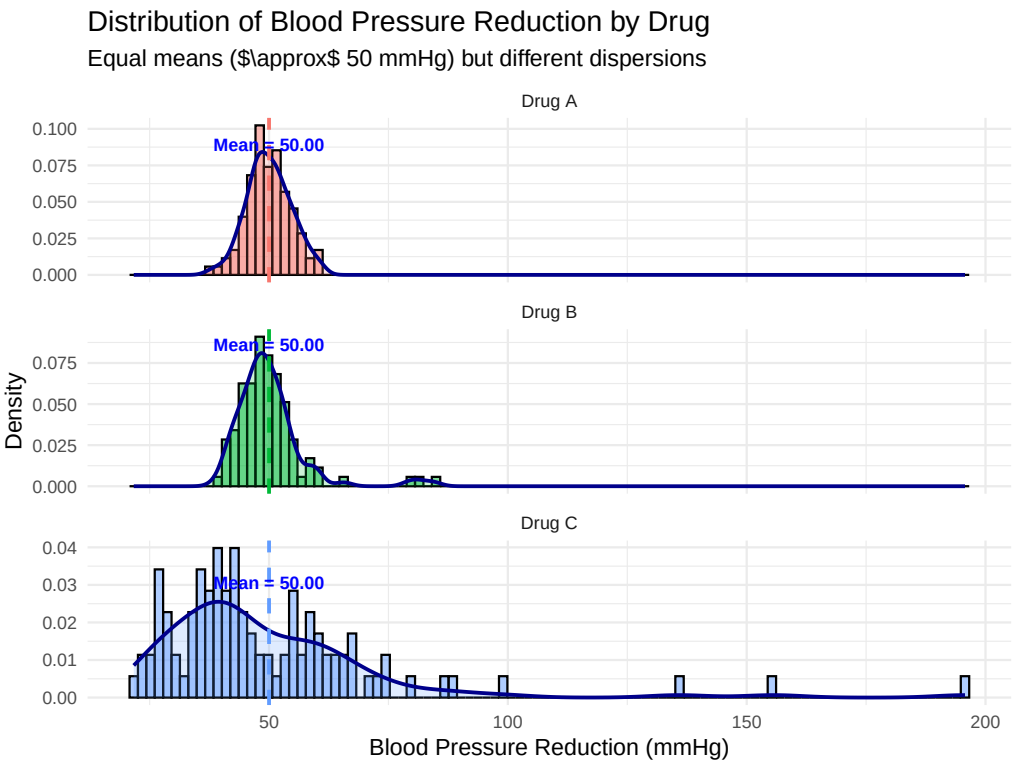
- **Drug A:** Symmetrical violin and narrow boxplot indicate low variability and a consistent effect among patients.
- **Drug B:** Violin shows slight widening at higher values; boxplot highlights mild outliers. Indicates moderate variability; most patients respond similarly, but a few have stronger effects.
- **Drug C:** Right-skewed violin with long tail and extreme points. Boxplot captures these extremes, showing high variability and skewness. Some patients experience much higher reductions than the majority.

### 5.4.2 Histograms

A histogram provides a visual summary of a dataset by dividing the range of data into consecutive intervals, called **bins**, and displaying the frequency or density of observations in each

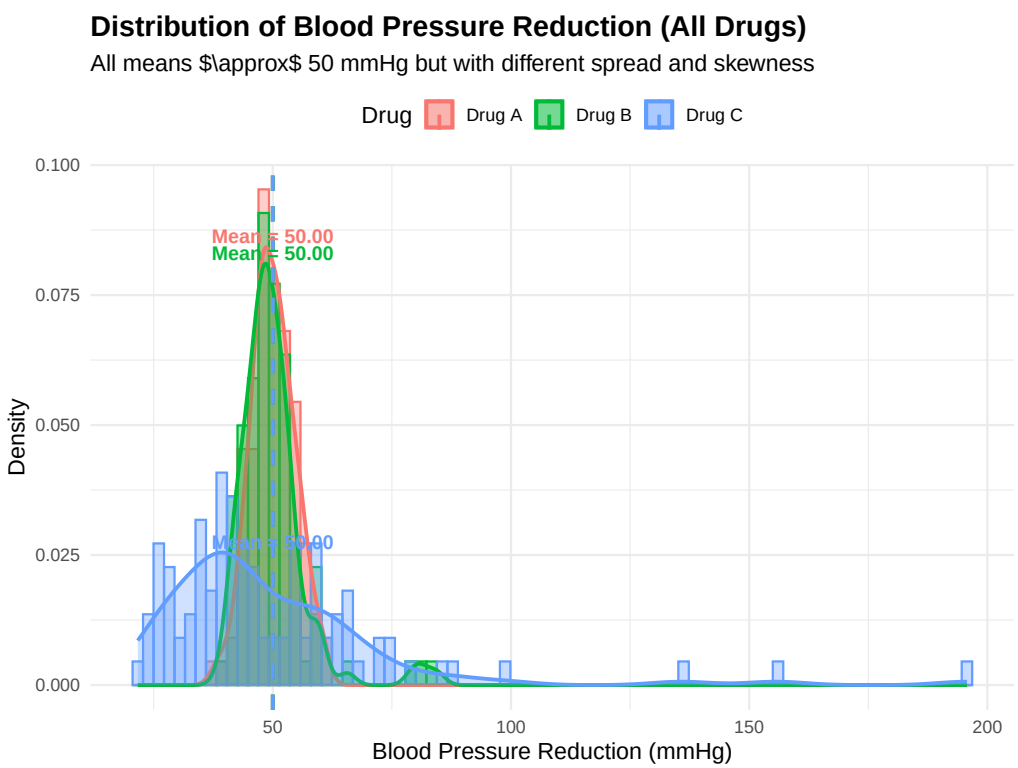
bin. The height of each bar reflects how many data points fall within that interval. Histograms allow us to quickly assess the shape of the distribution, the spread of the data, the presence of skewness, and potential outliers.

```
# -------- Plot: Histogram + Density + Smart Mean Label Placement --------
library(dplyr)
library(ggplot2)

# Calculate density peaks (for label positioning)
density_peaks <- drug_data %>%
  group_by(Drug) %>%
  summarise(PeakY = max(density(BP_Reduction)$y))

# Combine with mean values
label_data <- left_join(drug_summary, density_peaks, by = "Drug")

# Plot
ggplot(drug_data, aes(x = BP_Reduction, fill = Drug)) +
  geom_histogram(aes(y = after_stat(density)), alpha = 0.5,
                 color = "black", bins = 100, position = "identity") +
  geom_density(alpha = 0.2, color = "darkblue", size = 1) +
  geom_vline(data = drug_summary, aes(xintercept = Mean, color = Drug),
             linetype = "dashed", size = 1) +
  geom_text(
    data = label_data,
    aes(x = Mean, y = PeakY + 0.005,  # just above peak
        label = paste0("Mean = ", formatC(Mean, digits = 2, format = "f"))),
    color = "blue", size = 3.5, fontface = "bold"
  ) +
  facet_wrap(~Drug, ncol = 1, scales = "free_y") +
  labs(
    title = "Distribution of Blood Pressure Reduction by Drug",
    subtitle = "Equal means ($\\approx$ 50 mmHg) but different dispersions",
    x = "Blood Pressure Reduction (mmHg)",
    y = "Density"
  ) +
  theme_minimal(base_size = 13) +
  theme(legend.position = "none")
```




```
# -------- Plot: All Drugs in One Frame --------
library(dplyr)
library(ggplot2)

# Gabungkan mean + density peak (optional: tidak wajib kalau mau manual y posisi)
density_peaks <- drug_data %>%
  group_by(Drug) %>%
  summarise(PeakY = max(density(BP_Reduction)$y))

label_data <- left_join(drug_summary, density_peaks, by = "Drug")

ggplot(drug_data, aes(x = BP_Reduction, fill = Drug, color = Drug)) +
  # Histogram density-normalized
  geom_histogram(aes(y = after_stat(density)),
                 position = "identity", bins = 80, alpha = 0.35) +
  # Density curve
  geom_density(alpha = 0.3, linewidth = 1) +
  # Mean line
  geom_vline(data = drug_summary,
             aes(xintercept = Mean, color = Drug),
             linetype = "dashed", linewidth = 1) +
  # Mean label above each density peak
  geom_text(
    data = label_data,
    aes(x = Mean, y = PeakY + 0.002,
```

```
      label = paste0("Mean = ", formatC(Mean, digits = 2, format = "f")),
      color = Drug),
  size = 4, fontface = "bold", show.legend = FALSE
) +
labs(
  title = "Distribution of Blood Pressure Reduction (All Drugs)",
  subtitle = "All means $\\approx$ 50 mmHg but with different spread and skewness",
  x = "Blood Pressure Reduction (mmHg)",
  y = "Density",
  fill = "Drug",
  color = "Drug"
) +
theme_minimal(base_size = 13) +
theme(
  legend.position = "top",
  plot.title = element_text(face = "bold")
)
```




**Interpretation:**

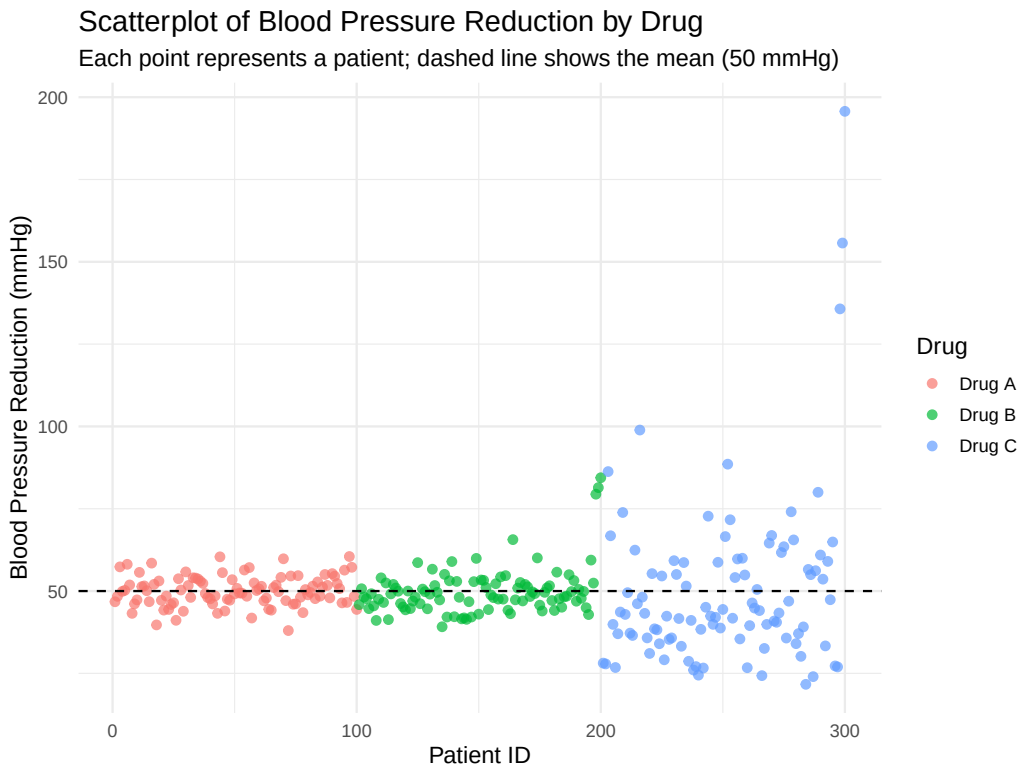
- **Drug A:** Narrow histogram and density curve indicate tight clustering around the mean. Low variability → consistent effect among patients.
- **Drug B:** Slightly wider spread and presence of minor outliers. Moderate variability → most patients respond similarly, but a few show extreme reduction.

- **Drug C:** Right-skewed distribution with long tail and extreme outliers. High variability → responses vary greatly, and some patients experience very high reductions.

### 5.4.3 Scatterplots

Here's a practical scatterplot example using Drug A, B, and C, showing blood pressure reductions for individual patients, including trend lines and mean reference.

```
# -------- Scatterplot --------
ggplot(drug_data, aes(x = PatientID, y = BP_Reduction, color = Drug)) +
  geom_point(size = 2, alpha = 0.7) +
  geom_hline(yintercept = 50, linetype = "dashed", color = "black") +
  labs(
    title = "Scatterplot of Blood Pressure Reduction by Drug",
    subtitle = "Each point represents a patient; dashed line shows the mean (50 mmHg)",
    x = "Patient ID",
    y = "Blood Pressure Reduction (mmHg)"
  ) +
  theme_minimal(base_size = 13)
```




**Explanation:

- Each point represents a patient's reduction in blood pressure.
- The x-axis shows individual patients, while the y-axis shows the BP reduction.

- The dashed line at 50 mmHg represents the mean reduction for all drugs.
- Drug A shows tightly clustered points (low variability).
- Drug B has some extreme points (outliers), increasing variability.
- Drug C shows a right-skewed distribution with extreme outliers, indicating high variability.

This scatterplot visually complements the histogram and boxplot analyses, helping to identify patient-level differences and treatment consistency.

# References

# Chapter 6

# Essentials of Probability

Probability is a foundational pillar of statistical reasoning, offering a systematic and coherent framework for understanding uncertainty and guiding informed decision-making. Rather than relying on intuition or conjecture, probability enables us to quantify the likelihood of various outcomes, interpret patterns within data, and analyze phenomena that arise from natural or experimental processes. A strong command of probability concepts is essential for effective data analysis, scientific research, and evidence-based practice.

This section presents the key principles that form the basis of probability theory:

- Fundamental concepts of probability, including sample spaces, events, and the complement rule—core components that define how probabilities are structured and interpreted.
- Independent and dependent events, which differentiate scenarios where the occurrence of one event does or does not influence another, a distinction critical for accurate modeling and prediction.
- The union of events, which addresses the probability that at least one among several events will occur.
- Exclusive and exhaustive events, clarifying how events interact within a sample space and how those relationships shape probability calculations.
- Binomial experiments and binomial distributions, essential tools for analyzing repeated trials with two possible outcomes, widely used in scientific studies, reliability testing, and survey analysis.

Each topic is accompanied by instructional video resources designed to enhance conceptual understanding and support deeper engagement with the material. Together, these components provide a comprehensive and rigorous foundation for advancing to more complex statistical methods.

## 6.1 Fundamental Concept

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/ynjHKBCiGXY?si=lmDmcVp-f1l64TbV
```

## 6.2 Independent and Dependent

Video cannot be displayed in PDF/Word.
Please view the HTML version or open directly on YouTube:
https://www.youtube.com/embed/LS-_ihDKr2M?si=HQq5ACfh5wwDYmiU

## 6.3 Union of Events

Video cannot be displayed in PDF/Word.
Please view the HTML version or open directly on YouTube:
https://www.youtube.com/embed/vqKAbhCqSTc?si=d3US5PYLeV-DZWBZ

## 6.4 Exclusive and Exhaustive

Video cannot be displayed in PDF/Word.
Please view the HTML version or open directly on YouTube:
https://www.youtube.com/embed/f7agTv9nA5k?si=SoKGr0XpHj1u4K5f

## 6.5 Binomial Experiment

Video cannot be displayed in PDF/Word.
Please view the HTML version or open directly on YouTube:
https://www.youtube.com/embed/nRuQAtajJYk?si=Q0Ulh5UxDF0svzXd

## 6.6 Binomial Distribution

Video cannot be displayed in PDF/Word.
Please view the HTML version or open directly on YouTube:
https://www.youtube.com/embed/Y2-vSWFmgyI?si=Tz7vLscgShvKrBxQ

## References

# Chapter 7

# Probability Distributions

Probability not only helps us understand how likely an event is to occur, but also forms the foundation of many statistical methods used for decision-making. When a process or experiment produces varying outcomes, we use a random variable to represent those outcomes and a probability distribution to describe how the probabilities are assigned to each possible value. Understanding the shape and properties of a distribution is essential because it determines how data behave, how we calculate probabilities, and how we make predictions. From distributions for continuous variables to the behavior of statistics such as sample means, probability distributions serve as the core of inferential statistics.

This material will guide you through several key concepts:

- Continuous Random Variables for continuous variables, which describe the likelihood of values over a continuous range.
- Sampling distributions, which represent the distribution of sample statistics such as the sample mean or sample proportion.
- The Central Limit Theorem (CLT), one of the most important results in statistics, explaining why the distribution of sample means tends to be normal regardless of the population's underlying shape.
- Sample proportion distributions, widely used in survey analysis and quantitative research.

Each section is supported with video explanations to enhance conceptual understanding. By mastering these topics, you will be better equipped to analyze data, build statistical models, and draw conclusions based on solid probabilistic principles.

## 7.1 Continuous Random

Understanding these basics will provide a strong foundation as we transition into the main topic of this video: Continuous Random Variables and Their Probability Distributions.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/ZyUzRVa6hCM?si=X5gzd8qqSrIIbtiI
```

To understand continuous random variables, it is essential to know how probability is represented using a **Probability Density Function (PDF)**.
Unlike discrete random variables, a continuous random variable does not assign probability to individual points. Instead, probability is obtained from the **area under the PDF curve**.

### 7.1.1 Random Variable

A random variable is **continuous** if it can take any value within an interval on the real number line.
Examples include: height, time, temperature, age, pressure, and velocity.

**Key characteristics:**

- The variable takes values in an interval such as $(a, b)$ or even $(-\infty, +\infty)$.
- The probability of any single point is always zero:

$$P(X = x) = 0$$

- Probabilities are meaningful only over intervals:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

### 7.1.2 Probability Density Funct.

A function $f(x)$ is a valid Probability Density Function (PDF) if it satisfies:

**1. Non-negativity**

$$f(x) \geq 0 \quad \forall x$$

**2. Total Area Equals 1**

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

**Interpretation:**

- Larger values of $f(x)$ indicate higher probability *density* around that value.
- However, $f(x)$ **is not a probability**; probabilities come from the area under the curve.

**Example PDF:** $f(x) = 3x^2$ on $[0, 1]$

Consider the probability density function:

$$f(x) = 3x^2, \quad 0 \leq x \leq 1$$

Validation:

$$\int_0^1 3x^2 \, dx = 1$$

### 7.1.3 Probability on an Interval

To compute probability within an interval:

$$P(a \leq X \leq b) = \int_a^b 3x^2 \, dx$$

Example:

$$P(0.5 \leq X \leq 1)$$

### 7.1.4 Cumulative Distribution Funct.

The Cumulative Distribution Function (CDF) is defined as:

$$F(x) = P(X \leq x) = \int_0^x 3t^2 \, dt = x^3$$

Relationship between PDF and CDF:

$$f(x) = F'(x)$$

## 7.2 Sampling Distributions

Before exploring the concept of sampling distributions in detail, this video provides a clear visual explanation of how statistics such as sample means behave when repeatedly drawn from the same population. It offers an intuitive foundation for understanding variability, uncertainty, and why sampling distributions are essential in statistical inference. Please watch the video below before continuing with the material.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/7S7j75d3GM4?si=8-iAi1t3dy13AgZL
```

## 7.3 Central Limit Theorem

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/ivd8wEHnMCg?si=EgHT8gPNfz13gwHR
```

## 7.4 Sample Proportion

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/q2e4mK0FTbw?si=36BHRKnztM-yXmPE
```

## 7.5 Review Sampling Distribution

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/cOmFEL_SWzE?si=diLTtIJ0cp-zVuz4
```

## References

# Chapter 8

# Confidence Interval

Before diving into the formulas and theory of **Confidence Intervals (CI)**, this chapter presents a video designed to help you grasp the concept in a simple and visual way. The video offers a clear picture of why CI matter and how they operate in data analysis. By understanding the core idea first, readers will be better prepared for the more detailed explanations that follow.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/MbXThbTSrVI?si=BZQL6DzC8ScRkb2b
```

The video above provides a visual and intuitive introduction to the fundamental ideas of inferential statistics—especially uncertainty, estimation, and significance in data analysis. This overview serves as a systematic guide to understand how each chapter connects before you explore them in detail. The framework and methods align with recent standard references such as [56], [57], and [58], which provide updated discussions on probability distributions, sampling theory, and CI construction and interpretation.

## 8.1 CI using z-Distribution

When estimating a population mean and the population standard deviation is known, or when the sample size is large (typically $n \geq 30$), we can use the normal ($z$) distribution to construct a Confidence Interval. The $z$-distribution (standard normal) has fixed variance, unlike the $t$-distribution whose variance depends on sample size.

Because of this, the $z$-distribution is appropriate when the variability of the population is already known or well-estimated from big data.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/czdwHU27OqA?si=1R-DCvlQDo-ACEWL
```

### 8.1.1 Manual of z-distribution

The analytics team wants to measure the **average number of daily clicks** on a new application feature. It is known that the **population standard deviation is already known**, because the historical dataset is very large. From the initial test, **50 user samples** tried the new feature, and the following summary was obtained:

**Summary data:**

- Sample mean:$\bar{x} = 23.8$ clicks/day
- Population standard deviation (known):$\sigma = 4.5$ clicks
- Sample size:$n = 50$

We want to calculate the **95% Confidence Interval** for the population mean of daily clicks.

**Formula for CI using *z-distribution***

$$CI = \bar{x} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$$

**Sample size**

$$n = 50$$

Because $n \geq 30$, the *z-distribution* is valid even if the population distribution is not known.

**Sample mean**

$$\begin{aligned} \bar{x} &= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} x_i \\ &= \tfrac{1}{50}(x_1 + x_2 + \cdots + x_n) \\ &= 23.8 \end{aligned}$$

**Population standard deviation (known)**

$$\sigma = 4.5$$

This is the main requirement for using the *z-distribution*.

**Critical value $z$ for 95% CI**

- Significance level: $\alpha = 0.05, \quad \alpha/2 = 0.025$
- Standard normal distribution table: $z_{0.025} = 1.96$

**Standard Error (SE)**

$$\begin{aligned} SE &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{4.5}{\sqrt{50}} \\ &= \frac{4.5}{7.071} \\ &\approx 0.637 \end{aligned}$$

**Margin of Error (ME)**

$$\begin{aligned} ME &= z_{\alpha/2} \times SE \\ &= 1.96 \times 0.637 \\ &\approx 1.248 \end{aligned}$$

**Confidence Interval**

$$\begin{aligned} CI_{95\%} &= \bar{x} \pm ME \\ &= 23.8 \pm 1.248 \\ &\approx (22.552,\ 25.048) \end{aligned}$$

**Interpretation (Data Science)**

With 95% confidence, the average number of daily clicks from users of the new feature is estimated to lie between:

$$22.552 \text{ and } 25.048 \text{ clicks per day}$$

The confidence interval from the *z-distribution* is **narrower** than that from the *t-distribution* because $\sigma$ is known and does not need to be estimated from the sample.

### 8.1.2 R Code for z-distribution

```r
# Load libraries
library(knitr)
library(kableExtra)
library(htmltools)

# Data input
xbar <- 23.8                  # sample mean
sigma <- 4.5                  # population standard deviation (known)
n <- 50                       # sample size
alpha <- 0.05                 # significance level
z_crit <- qnorm(1 - alpha/2) # Critical z-value for 95% CI
SE <- sigma / sqrt(n)         # Standard Error (SE)
ME <- z_crit * SE             # Margin of Error (ME)
lower_CI <- xbar - ME         # LCI
```

```r
upper_CI <- xbar + ME        # UCI

# Summary table with formulas (LaTeX)
summary_table <- data.frame(
  Parameter = c("Sample mean (x̄)",
                "Population SD ( )",
                "Sample size (n)",
                "z critical value",
                "Standard Error (SE)",
                "Margin of Error (ME)",
                "Lower CI",
                "Upper CI"),
  Value = c(xbar, sigma, n, round(z_crit,4),
            round(SE,4), round(ME,4), round(lower_CI,3), round(upper_CI,3)),
  Formula = c(
    "$$\\bar{x} = \\frac{1}{n}\\sum_{i=1}^{n} x_i$$",
    "$$\\sigma$$",
    "$$n$$",
    "$$z_{1-\\alpha/2}$$",
    "$$SE = \\frac{\\sigma}{\\sqrt{n}}$$",
    "$$ME = z_{1-\\alpha/2} \\times SE$$",
    "$$\\bar{x} - ME$$",
    "$$\\bar{x} + ME$$"
  ),
  stringsAsFactors = FALSE
)

# Render tabel in Quarto HTML
kable(summary_table, escape = FALSE, booktabs = TRUE, align = "lcc") %>%
  kable_styling(full_width = FALSE)
```

| Parameter | Value | Formula |
|---|---|---|
| Sample mean (x̄) | 23.8000 | $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ |
| Population SD (σ) | 4.5000 | $$\sigma$$ |
| Sample size (n) | 50.0000 | $$n$$ |
| z critical value | 1.9600 | $$z_{1-\alpha/2}$$ |
| Standard Error (SE) | 0.6364 | $$SE = \frac{\sigma}{\sqrt{n}}$$ |
| Margin of Error (ME) | 1.2473 | $$ME = z_{1-\alpha/2} \times SE$$ |
| Lower CI | 22.5530 | $$\bar{x} - ME$$ |
| Upper CI | 25.0470 | $$\bar{x} + ME$$ |

## 8.2 CI Using t-Distribution

When estimating a population mean from sample data, we often do **not** know the true population standard deviation. In these situations—especially when the sample size is small—the

**t-distribution** provides a more accurate way to measure uncertainty than the normal ($z$) distribution.

The $t$-distribution has **heavier tails**, reflecting the extra variability that comes from estimating the standard deviation directly from the sample.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/6r5IZCjvIHI?si=9qi2oF7GSrzlMZit
```

A **Confidence Interval (CI)** for the population mean using the $t$-distribution is:

$$\bar{x} \pm t_{\alpha/2,\, df} \left( \frac{s}{\sqrt{n}} \right)$$

where:

- $\bar{x}$ = sample mean
- $s$ = sample standard deviation
- $n$ = sample size
- $df = n - 1$ = degrees of freedom
- $t_{\alpha/2,\, df}$ = critical $t$-value from the $t$-distribution

Using this formula, we create an interval that likely contains the true population mean, while realistically accounting for uncertainty due to limited data.

### 8.2.1 Manual of t-distribution

The product team launched a new recommendation feature and took a small sample of user interactions to measure *engagement*, specifically the **time spent (in minutes)** on the feature. We want to estimate the average time spent by all users on this feature with a 95% confidence level.

**Sample data (minutes):** $7.2,\ 5.8,\ 6.5,\ 8.0,\ 6.9,\ 7.4,\ 5.5,\ 6.7,\ 7.1,\ 6.3$

**Sample size**

$$n = 10$$

**Sample mean**

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{10}(x_1 + x_2 + \cdots + x_{10}) \\ &= \frac{1}{10}(7.2 + 5.8 + 6.5 + 8.0 + 6.9 + 7.4 + 5.5 + 6.7 + 7.1 + 6.3) \\ &= \frac{67.4}{10} \\ &= 6.74 \text{ minutes}\end{aligned}$$

#### Sample standard deviation {-}

$$\begin{aligned}s &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{(7.2-6.74)^2 + (5.8-6.74)^2 + \cdots + (6.3-6.74)^2}{10-1}} \\ &= \sqrt{\frac{4.063}{9}} \\ &\approx 0.67 \text{ minutes}\end{aligned}$$

**Degrees of freedom**

$$df = n - 1 = 9$$

**Critical $t$ value for 95% CI**

Significance level:

$$\alpha = 0.05, \quad \alpha/2 = 0.025$$

From the t-table (or statistical function):

$$t_{0.025,\,9} \approx 2.262$$

**Standard Error (SE)**

$$\begin{aligned}SE &= \frac{s}{\sqrt{n}} \\ &= \frac{0.77}{\sqrt{10}} \\ &\approx 0.2436\end{aligned}$$

**Margin of Error (ME)**

$$\begin{aligned}ME &= t_{\alpha/2,\,df} \times SE \\ &= 2.262 \times 0.2436 \\ &\approx 0.551\end{aligned}$$

**Confidence Interval**

$$\begin{aligned} CI_{95\%} &= \bar{x} \pm ME \\ &= 6.74 \pm 0.551 \\ &\approx (6.203,\ 7.291) \end{aligned}$$

**Interpretation (Data Science)**

With 95% confidence, we estimate that the **average time spent by users on the recommendation feature is between 6.203 and 7.291 minutes**. Even with a small sample, this interval provides a realistic range for the population mean, accounting for uncertainty from estimating the standard deviation.

### 8.2.2 R Code t-distribution

```r
library(knitr)
library(kableExtra)

# Data
data <- c(7.2,5.8,6.5,8.0,6.9,7.4,5.5,6.7,7.1,6.3)
n <- length(data)
xbar <- mean(data)
s <- sd(data)
df <- n - 1
alpha <- 0.05
t_crit <- qt(1 - alpha/2, df)
SE <- s / sqrt(n)
ME <- t_crit * SE
CI_lower <- xbar - ME
CI_upper <- xbar + ME

# Summary table with formulas
summary_table <- data.frame(
  Parameter = c("Sample size (n)",
                "Sample mean (x̄)",
                "Sample SD (s)",
                "Degrees of freedom (df)",
                "t critical value",
                "Standard Error (SE)",
                "Margin of Error (ME)",
                "Lower CI",
                "Upper CI"),
  Value = c(n, round(xbar,3),
            round(s,3), df, round(t_crit,3),
            round(SE,3), round(ME,3),
            round(CI_lower,3), round(CI_upper,3)),
  Formula = c(
```

```
    "$$n$$",
    "$$\\bar{x} = \\frac{1}{n}\\sum_{i=1}^{n} x_i$$",
    "$$s = \\sqrt{\\frac{\\sum_{i=1}^{n} (x_i - \\bar{x})^2}{n-1}}$$",
    "$$df = n-1$$",
    "$$t_{1-\\alpha/2, df}$$",
    "$$SE = \\frac{s}{\\sqrt{n}}$$",
    "$$ME = t_{1-\\alpha/2} \\times SE$$",
    "$$\\bar{x} - ME$$",
    "$$\\bar{x} + ME$$"
  ),
  stringsAsFactors = FALSE
)

# Render table
kable(summary_table, escape = FALSE, booktabs = TRUE, align = "lcc") %>%
  kable_styling(full_width = FALSE)
```

| Parameter | Value | Formula |
|---|---|---|
| Sample size (n) | 10.000 | $$n$$ |
| Sample mean ($\bar{x}$) | \| 6.740 | \| $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ |
| Sample SD (s) | 0.750 | $$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$ |
| Degrees of freedom (df) | 9.000 | $$df = n-1$$ |
| t critical value | 2.262 | $$t_{1-\alpha/2, df}$$ |
| Standard Error (SE) | 0.237 | $$SE = \frac{s}{\sqrt{n}}$$ |
| Margin of Error (ME) | 0.537 | $$ME = t_{1-\alpha/2} \times SE$$ |
| Lower CI | 6.203 | $$\bar{x} - ME$$ |
| Upper CI | 7.277 | $$\bar{x} + ME$$ |

## 8.3 Determining the Sample Size

Determining the **sample size** is a crucial step in designing experiments, surveys, and data analyses. The goal is to ensure that the sample is large enough to provide **accurate and reliable estimates** of population parameters such as the mean $\mu$ or proportion $p$. Sample size calculations typically depend on:

- confidence level
- acceptable margin of error
- variability in the data (e.g., standard deviation)
- whether the population is large or finite

When the population standard deviation $\sigma$ is known, the minimum required sample size is:

$$n = \left( \frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

where:

- $z_{\alpha/2}$ = critical value from the standard normal distribution
- $\sigma$ = population standard deviation
- $E$ = desired margin of error

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/qVDVAZigXg0?si=GbqdAjaSclnitViq
```

### 8.3.1 Manual of the Sample Size

A data analytics team wants to estimate the **average page loading time** in an application. From historical data, the population standard deviation is known to be:

$$\sigma = 1.8 \text{ seconds}$$

The team wants a margin of error of:

$$E = 0.3 \text{ seconds}$$

Confidence level:

$$95\%, \quad z_{0.025} = 1.96$$

$$n = \left( \frac{1.96 \times 1.8}{0.3} \right)^2$$

$$\begin{aligned} n &= \left(\tfrac{3.528}{0.3}\right)^2 \\ &= (11.76)^2 \\ &\approx 138.3 \end{aligned}$$

Sample size must be an integer:

$$n = 139 \text{ observations}$$

### 8.3.2 R Code (Sample Size for a Mean)

```r
sigma <- 1.8
E <- 0.3
z <- 1.96

n <- (z * sigma / E)^2
ceiling(n)
```

```
[1] 139
```

## 8.4 CI for a Proportion

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/dLEtlteLVJU?si=82teF9pD7rZV_6B8
```

### 8.4.1 Manual of CI Proportion

A data science team wants to estimate the **proportion of users who clicked** on a new call-to-action (CTA) button during an A/B test. From a sample of users, the team records how many actually clicked the button.

#### Sample Data

- Total sample size: $n = 240$,
- Number of users who clicked: $x = 78$

#### Sample Proportion

The sample proportion $\hat{p}$ is:

$$\hat{p} = \frac{x}{n} = \frac{78}{240} = 0.325$$

So about **32.5%** of sampled users clicked the CTA. Compute a **95% confidence interval** for the true population proportion $p$.

#### Standard Error (SE)

$$\begin{aligned} SE &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= \sqrt{\frac{0.325(1-0.325)}{240}} \\ &= \sqrt{\frac{0.325 \times 0.675}{240}} \\ &\approx 0.0294 \end{aligned}$$

**Critical Value**

For a 95% confidence level: $z_{\alpha/2} = 1.96$

**Margin of Error (ME)**

$$ME = z_{\alpha/2} \times SE = 1.96 \times 0.0294 \approx 0.0577$$

**Confidence Interval**

$$\begin{aligned} CI_{95\%} &= \hat{p} \pm ME \\ &= 0.325 \pm 0.0577 \\ &\approx (0.267,\ 0.383) \end{aligned}$$

**Interpretation (Data Science)**

With 95% confidence, the **true proportion of all users who would click the CTA lies between 26.7% and 38.3%**. This interval quantifies uncertainty and helps the team decide whether the CTA is performing strongly enough for deployment.

### 8.4.2 R Code for CI Proportion

```
library(knitr)
library(kableExtra)

# Data
n <- 240
x <- 78
p_hat <- x / n
z_crit <- 1.96
SE <- sqrt(p_hat * (1 - p_hat) / n)
ME <- z_crit * SE
CI_lower <- p_hat - ME
CI_upper <- p_hat + ME

# Summary table with formulas
summary_table <- data.frame(
  Parameter = c("Total sample size (n)",
                "Number of successes (x)",
                "Sample proportion (p̂)",
                "Standard Error (SE)",
                "Critical value (z_{ /2})",
                "Margin of Error (ME)",
                "Lower 95% CI",
                "Upper 95% CI"),
```

```
  Value = c(n, x, round(p_hat,3), round(SE,4), z_crit, round(ME,4), round(CI_lower,3), rou
  Formula = c(
    "$$n$$",
    "$$x$$",
    "$$\\hat{p} = \\frac{x}{n}$$",
    "$$SE = \\sqrt{\\frac{\\hat{p}(1-\\hat{p})}{n}}$$",
    "$$z_{1-\\alpha/2}$$",
    "$$ME = z_{1-\\alpha/2} \\times SE$$",
    "$$\\hat{p} - ME$$",
    "$$\\hat{p} + ME$$"
  ),
  stringsAsFactors = FALSE
)

# Render table
kable(summary_table, escape = FALSE, booktabs = TRUE, align = "lcc") %>%
  kable_styling(full_width = FALSE)
```

| Parameter | Value | Formula |
|---|---|---|
| Total sample size (n) | 240.0000 | $$n$$ |
| Number of successes (x) | 78.0000 | $$x$$ |
| Sample proportion (p̂) | \| 0.3250 | \| $$\hat{p} = \frac{x}{n}$$ |
| Standard Error (SE) | 0.0302 | $$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ |
| Critical value (z_{α/2}) | 1.9600 | $$z_{1-\alpha/2}$$ |
| Margin of Error (ME) | 0.0593 | $$ME = z_{1-\alpha/2} \times SE$$ |
| Lower 95% CI | 0.2660 | $$\hat{p} - ME$$ |
| Upper 95% CI | 0.3840 | $$\hat{p} + ME$$ |

## 8.5 One-Sided CI

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/c9RVFq6v5-g?si=Bdc9-BPQO4L1Va1d
```

### 8.5.1 Manual of One-Sided CI

The product team launched a new recommendation feature and sampled user interactions to measure **engagement**, specifically **the proportion of users who clicked the CTA**. We want to estimate the **true population proportion** with a **95% one-sided confidence interval**.

**Sample data:**

- Total sample size: $n = 240$
- Number of users who clicked: $x = 78$

**Sample proportion**

$$\hat{p} = \frac{x}{n} = \frac{78}{240} = 0.325$$

**Standard Error (SE)**

$$\begin{aligned} SE &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= \sqrt{\frac{0.325 \times 0.675}{240}} \\ &\approx 0.0294 \end{aligned}$$

**Critical value for 95% one-sided CI**

Significance level:

$$\alpha = 0.05$$

From z-table (one-sided):

$$z_{1-\alpha} \approx 1.645$$

**Margin of Error (ME)**

$$\begin{aligned} ME &= z_{1-\alpha} \cdot SE \\ &= 1.645 \cdot 0.0294 \\ &\approx 0.0484 \end{aligned}$$

**One-Sided Confidence Interval**

**Upper One-Sided CI:**

$$\begin{aligned} CI_{upper} &= \hat{p} + ME \\ &= 0.325 + 0.0484 \\ &\approx 0.373 \end{aligned}$$

**Lower One-Sided CI:**

$$\begin{aligned} CI_{lower} &= \hat{p} - ME \\ &= 0.325 - 0.0484 \\ &\approx 0.277 \end{aligned}$$

**Interpretation (Data Science)**

- **Lower one-sided CI:** With 95% confidence, at least **27.7%** of users would click the CTA.
- **Upper one-sided CI:** With 95% confidence, no more than **37.3%** of users would click the CTA.

This interval quantifies uncertainty in the population proportion using **one-sided estimation**, which is useful for decision-making when we are only concerned with a **minimum** or **maximum** threshold.

### 8.5.2 R Code One-Sided CI

```
library(knitr)
library(kableExtra)

# Data
n <- 240       # sample size
x <- 78        # number of successes
p_hat <- x/n   # sample proportion
alpha <- 0.05
z_crit <- qnorm(1 - alpha)   # one-sided z critical
SE <- sqrt(p_hat * (1 - p_hat)/n)
ME <- z_crit * SE
CI_lower <- p_hat - ME
CI_upper <- p_hat + ME

# Summary table with formulas
summary_table <- data.frame(
  Parameter = c("Sample size (n)",
                "Number of successes (x)",
                "Sample proportion (p̂)",
                "Significance level ( )",
                "Critical value (z  )",
                "Standard Error (SE)",
                "Margin of Error (ME)",
                "Lower One-Sided CI",
                "Upper One-Sided CI"),
  Value = c(n, x, round(p_hat,3), alpha, round(z_crit,3), round(SE,4), round(ME,4), round(
  Formula = c(
    "$$n$$",
    "$$x$$",
    "$$\\hat{p} = \\frac{x}{n}$$",
    "$$\\alpha$$",
    "$$z_{1-\\alpha}$$",
    "$$SE = \\sqrt{\\frac{\\hat{p}(1-\\hat{p})}{n}}$$",
    "$$ME = z_{1-\\alpha} \\times SE$$",
```

```
    "$$CI_{lower} = \\hat{p} - ME$$",
    "$$CI_{upper} = \\hat{p} + ME$$"
  ),
  stringsAsFactors = FALSE
)

# Render table
kable(summary_table, escape = FALSE, booktabs = TRUE, align = "lcc") %>%
  kable_styling(full_width = FALSE)
```

| Parameter | Value | Formula |
|---|---|---|
| Sample size (n) | 240.0000 | $$n$$ |
| Number of successes (x) | 78.0000 | $$x$$ |
| Sample proportion (p̂) | \| 0.3250 | \| $$\hat{p} = \frac{x}{n}$$ |
| Significance level (α) | 0.0500 | $$\alpha$$ |
| Critical value (z□□α) | 1.6450 | $$z_{1-\alpha}$$ |
| Standard Error (SE) | 0.0302 | $$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ |
| Margin of Error (ME) | 0.0497 | $$ME = z_{1-\alpha} \times SE$$ |
| Lower One-Sided CI | 0.2750 | $$CI_{lower} = \hat{p} - ME$$ |
| Upper One-Sided CI | 0.3750 | $$CI_{upper} = \hat{p} + ME$$ |

# References

# Chapter 9

# Statistical Inference

**Statistical inference** is the process of drawing conclusions about a population based on information obtained from a sample. It allows researchers and analysts to make generalizations, predictions, and decisions under uncertainty, bridging the gap between observed data and the broader population [59].

This mind map illustrates the core structure of statistical inference (see Figure Figure 9.1), highlighting its three main components: Statistical Hypotheses, Hypothesis Testing Methods, and Statistical Decision Making. Key elements such as Null Hypothesis (H□), Alternative Hypothesis (H□), T-Test, Z-Test, Chi-Square Test, and P-Value for decision making are included, providing a concise overview of how hypotheses are formulated, tested, and used to guide statistical decisions.

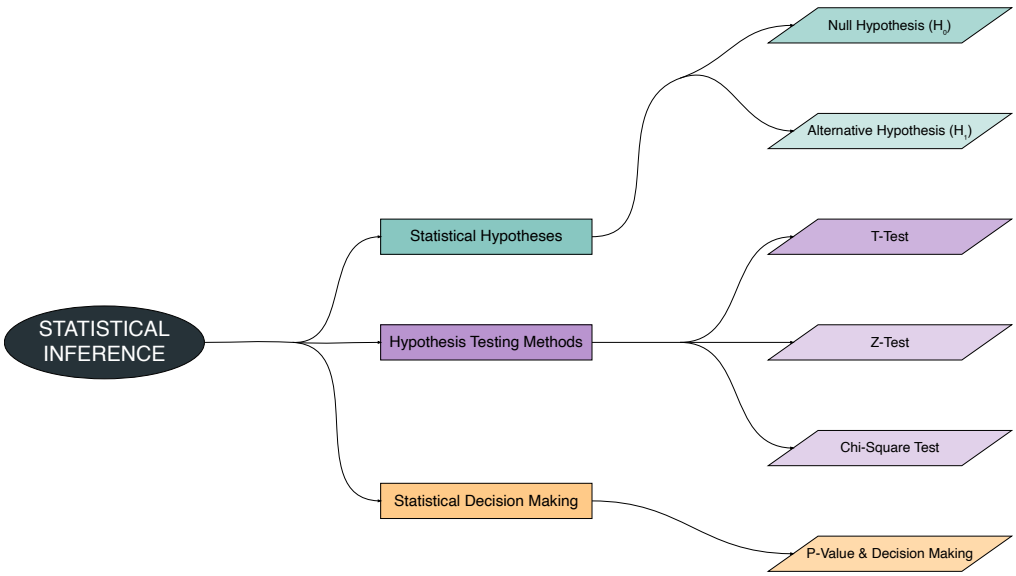



Figure 9.1: Statistical Inference

Statistical inference explains how conclusions about a population can be drawn from sample data by systematically accounting for uncertainty and variability. This topic serves as a bridge be-

tween descriptive statistics and formal decision-making methods, such as parameter estimation and hypothesis testing. The following video is designed to provide a clear and intuitive introduction to statistical inference, helping students build conceptual foundations that will support their learning of more advanced statistical techniques.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/6E6pB5JFLgM?si=t_lnDhsa80kJKA7k
```

## 9.1 Statistical Hypotheses

**Statistical hypotheses** are formal statements about a population parameter that can be tested using sample data. They provide a framework for making **objective decisions** based on evidence, helping researchers determine whether observed effects are due to random variation or represent a true phenomenon. In hypothesis testing, we compare the **Null Hypothesis (H□)** and the **Alternative Hypothesis (H□)** to decide which statement is more consistent with the observed data.

The following video provides a clear and concise conceptual explanation of the relationship between statistical hypotheses, parameter estimation, and confidence intervals, helping you build strong intuition before moving on to formal calculations and analytical procedures.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/a_l991xUAOU?si=rKXtI_DRUkgQkv8q
```

### 9.1.1 Null Hypothesis (H□)

The **Null Hypothesis (H□)** serves as the **baseline or reference point** in hypothesis testing. It represents the assumption that there is **no effect, no difference, or no relationship** in the population. H□ provides a standard against which the observed data is compared, allowing researchers to determine whether any observed difference is likely due to **random variation** rather than a true effect [60].

**Key Points:**

- Acts as a **benchmark** for testing statistical evidence.
- **Assumed true** initially and is **tested for possible rejection**, not proven.
- Denoted as **H□** in all statistical analyses.

**Examples:**

1. **Drug Effect on Blood Pressure:**
   The null hypothesis states that a new drug has **no effect** on blood pressure:

$$H_0 : \mu_{\text{treatment}} = \mu_{\text{control}}$$

2. **Comparison of Teaching Methods:**
   The null hypothesis states there is **no difference** in average test scores between two teaching methods:

$$H_0 : \mu_1 = \mu_2$$

In practice, H□ provides a **conservative assumption**. Only if the sample data provides **strong enough evidence** against H□ do we consider rejecting it in favor of the **Alternative Hypothesis (H□)**. This ensures decisions are **data-driven** and **objective**, minimizing the risk of concluding an effect exists when it does not [59].

### 9.1.2 Alternative Hypothesis (H□)

The **Alternative Hypothesis (H□ or Ha)** represents the statement that **contradicts the Null Hypothesis (H□)**. It reflects the effect, difference, or relationship the researcher expects to detect in the population [59]–[61].

**Key Points:**

- Indicates the presence of a **real effect or difference**.
- Denoted as **H□** or **Ha**.
- Can be **two-tailed** (detecting a difference in either direction) or **one-tailed** (detecting a difference in a specific direction).

**Examples:**

1. **Drug Effect on Blood Pressure:**
   The alternative hypothesis states that a new drug **reduces or changes blood pressure**:

$$H_1 : \mu_{\text{treatment}} \neq \mu_{\text{control}}$$

2. **Comparison of Teaching Methods:**
   The alternative hypothesis states that one teaching method **improves test scores** compared to the other:

$$H_1 : \mu_1 > \mu_2$$

In hypothesis testing, H□ is **accepted only if the sample evidence is strong enough to reject H□**. This ensures that conclusions about population effects are **supported by data**, reducing the risk of drawing incorrect inferences [59].

### 9.1.3 Type I/II Errors

In hypothesis testing, errors can occur when making decisions based on sample data. The two main types of errors are **Type I Error (α)** and **Type II Error (β)**. The table below summarizes the comparison with examples:

| Error Type | Definition | Probability | Example |
|---|---|---|---|
| **Type I Error (α)** | Rejecting H□ when it is actually true (false positive) | α, commonly 0.05 | Concluding a new drug lowers blood pressure when in reality it does not. |
| **Type II Error (β)** | Failing to reject H□ when it is actually false (false negative) | β; Power = 1 − β | Concluding a new drug has no effect on blood pressure when it actually does. |

**Notes:**

- **Type I Error (α)** is controlled by setting the **significance level** before conducting the test.
- **Type II Error (β)** depends on the **sample size, effect size, and variability**. Increasing sample size reduces β and increases the power of the test.
- Understanding both errors is crucial for **making informed statistical decisions** and balancing the risk of false positives and false negatives in research.

## 9.2 Hypothesis Test Methods

In statistics, **hypothesis testing methods** are used to determine whether the evidence from a sample is strong enough to reject the null hypothesis (H□) in favor of the alternative hypothesis (H□). The choice of test depends on the **type of data**, **sample size**, and **population characteristics**.

### 9.2.1 T-Test

The **T-Test** is used to compare the mean of a sample to a known value or to compare means between two groups when the **population standard deviation is unknown** and the sample size is relatively small.

**Types of T-Test:**

1. **One-sample T-Test:** Compare sample mean to a known value.

2. **Independent two-sample T-Test:** Compare means of two independent groups.

3. **Paired T-Test:** Compare means of **paired observations** (e.g., before-after measurements).

**Example:**
- Testing whether the average test score of students differs from 75.

$$H_0 : \mu = 75 \quad vs \quad H_1 : \mu \neq 75$$

### 9.2.2 Z-Test

The **Z-Test** is used to compare means when the **population standard deviation is known** or the sample size is **large (n ≥ 30)**. It assumes that the data is approximately normally distributed.

**Types of Z-Test:**

1. **One-sample Z-Test:** Compare a sample mean to a known population mean.

2. **Two-sample Z-Test:** Compare means of two independent populations with known standard deviations.

**Example:**

- Testing whether a new teaching method changes the average score, assuming the population standard deviation is known:

$$H_0 : \mu_{\text{new}} = \mu_{\text{old}} \quad vs \quad H_1 : \mu_{\text{new}} \neq \mu_{\text{old}}$$

### 9.2.3 Chi-Square Test

The **Chi-Square Test ($\chi^2$ Test)** is used for **categorical data** to examine whether the observed frequency distribution differs from the expected distribution.

**Types of Chi-Square Test:**

1. **Goodness-of-Fit Test:** Tests if a single categorical variable follows a hypothesized distribution.

2. **Test of Independence:** Tests whether two categorical variables are **independent**.

**Example:**

- Testing whether **gender (male/female)** is independent of **preference for online learning (yes/no)**:

$H_0$ : Gender and preference are independent$H_1$ : Gender and preference are not independent

## 9.3 Statistical Decision Making

**Statistical Decision Making** involves using the results of hypothesis tests to make **informed decisions** about the population. After performing a T-Test, Z-Test, or Chi-Square Test, we interpret the **p-value** and decide whether to **reject or fail to reject the null hypothesis (H□)**. This process allows us to draw conclusions while considering the **risk of errors**. Steps in Statistical Decision Making:

1. **Set significance level (α):**
    - Common choices: 0.05 (5%), 0.01 (1%), or 0.10 (10%).
    - This determines the threshold for rejecting H□.
2. **Perform the hypothesis test:**
    - Calculate test statistic $(T, Z, \chi^2)$ based on sample data.
    - Compute the **p-value**.
3. **Compare p-value with α:**
    - **If p-value ≤ α:** Reject H□ → evidence supports H□.
    - **If p-value > α:** Fail to reject H□ → insufficient evidence to support H□.
4. **Consider Type I and Type II Errors:**
    - Type I Error (α): Rejecting H□ when it is true.
    - Type II Error (β): Failing to reject H□ when it is false.
    - Balance between α and β is important for decision reliability.

### 9.3.1 T-Test

- Suppose we test whether the average score of a sample of students is different from 80.
- T-Test yields: **t = 1.82, p-value = 0.10**
- Significance level: α = 0.05

**Decision:**

- Since p-value (0.10) > α (0.05), we **fail to reject H□**.
- Interpretation: There is **insufficient evidence** to conclude that the average score differs from 80.

### 9.3.2 Chi-Square Test

- Suppose we test whether gender and preference for online learning are independent.
- Chi-Square Test yields: $\chi^2 = 4.23$, p-value = 0.04
- Significance level: $\alpha = 0.05$

**Decision:**

- Since p-value (0.04) < $\alpha$ (0.05), we **reject H□**.
- Interpretation: There is **evidence that gender and preference are not independent**.

# References

# Chapter 10

# Nonparametric Methods

**Nonparametric methods** are statistical techniques that do **not rely on strict distributional assumptions**, such as normality or known population parameters. These methods are particularly useful when dealing with **small samples**, **ordinal or categorical data**, or data that contain **outliers and skewness** [62]; [63].

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/welIUs2boQ8?si=9sKOrtKCAR8jMuAB
```

Within the framework of **statistical inference** (see, Figure 10.1), nonparametric methods allow researchers to draw valid conclusions even when classical parametric assumptions are violated. As a result, they are widely applied in **data science, business analytics, engineering, health sciences, and social research** [64].

## 10.1 Role of Nonparametric

Statistical inference aims to draw conclusions about a population based on sample data while accounting for uncertainty. Traditional parametric inference relies on assumptions about population parameters, such as the mean and variance. When these assumptions are questionable, **nonparametric inference provides a robust alternative** [62].

Instead of focusing on parameters like the mean ($\mu$), nonparametric methods often emphasize:

- Medians
- Distributional equality
- Ranks or signs
- Frequencies

This shift allows inference to remain valid under broader conditions.

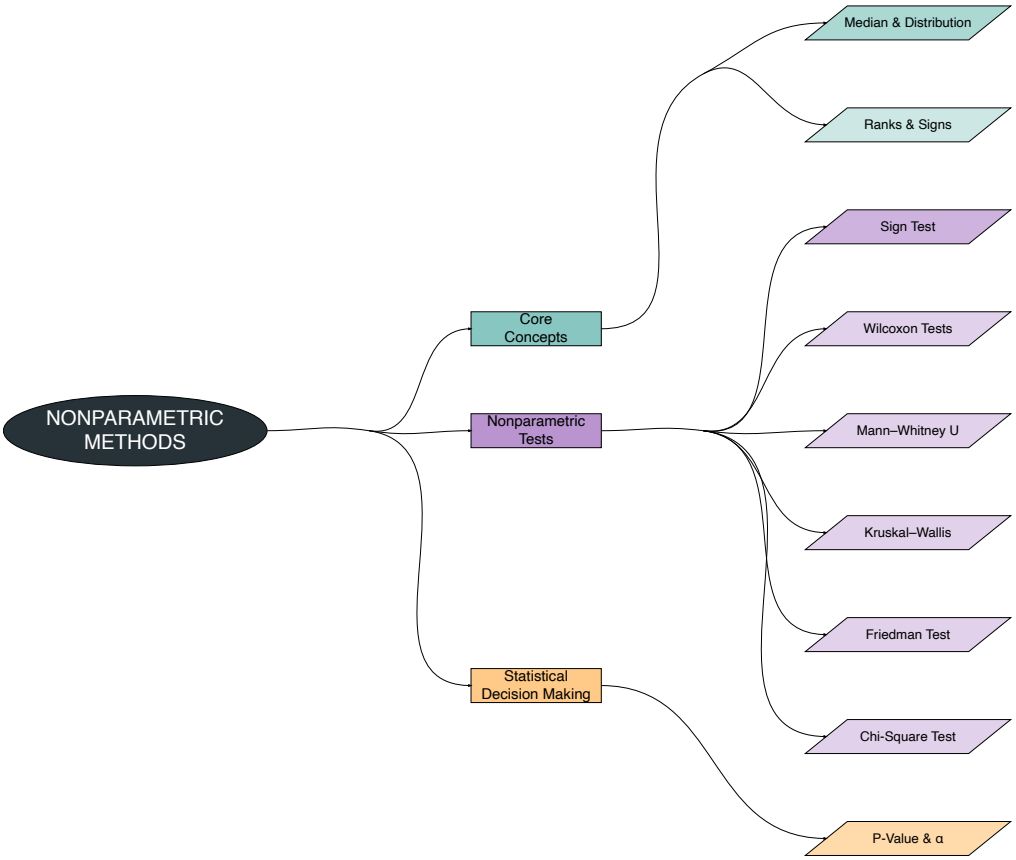



Figure 10.1: Nonparametric Methods

## 10.2 When to Use?

Nonparametric methods are recommended when:

- The data distribution is **unknown or non-normal**
- Sample size is **small**
- Data contain **outliers**
- Measurement scale is **ordinal or nominal**
- Variance homogeneity assumptions are violated

**Parametric vs Parametric Methods:**

| Aspect | Parametric Methods | Nonparametric Methods |
|---|---|---|
| Distribution assumption | Required | Not required |
| Sensitivity to outliers | High | Low |
| Data scale | Interval / Ratio | Ordinal / Nominal |
| Statistical power | Higher (if assumptions met) | Lower but more robust |

## 10.3 Nonparametric Hypotheses

As in parametric inference, nonparametric testing is based on **statistical hypotheses**:

- **Null Hypothesis (H□):** No difference, no effect, or no association
- **Alternative Hypothesis (H□):** A difference, effect, or association exists

However, these hypotheses typically concern **medians, distributions, or ranks**, rather than means [63].

**Example:**

$$H_0 : \text{Median satisfaction score is the same across services}$$

$$H_1 : \text{Median satisfaction score differs across services}$$

## 10.4 Common Nonparametric

### 10.4.1 Sign Test

The **Sign Test** is one of the simplest nonparametric tests and is used to test hypotheses about a population **median** or the median of paired differences. Unlike parametric tests, it does not require assumptions about the underlying data distribution and is highly robust to outliers.

**The Sign Test is appropriate when:**

- Observations are paired (before–after or matched samples)
- The distribution of differences is unknown or highly skewed
- Data contain outliers
- Only the direction of change is reliable

**Key Characteristics:**

- Uses only the sign (+ or −) of differences
- Extremely robust to non-normality and outliers
- Has relatively low statistical power

**Hypotheses**

$$H_0 : \text{Median difference} = 0$$

$$H_1 : \text{Median difference} \neq 0$$

**Real-World Case: Manufacturing Quality Control**

A manufacturing plant introduces a new **machine calibration procedure** aimed at reducing product defects. For each production batch, the number of defective items is recorded **before** and **after** calibration. Due to occasional machine failures, the defect counts include extreme values and are not normally distributed.

Because the magnitude of changes is unreliable, the **Sign Test** is applied to determine whether the **median change in defect counts** differs from zero.

**Step 1: Compute Paired Differences**

For each batch:

$$d_i = \text{Defects}_{\text{after}} - \text{Defects}_{\text{before}}$$

```
# Defect counts before and after calibration
before <- c(12, 15, 10, 18, 20, 14, 16, 22, 11, 19)
after  <- c(10, 14, 11, 15, 18, 13, 15, 20, 11, 17)

# Compute paired differences
diff <- after - before
diff
```

```
 [1] -2 -1  1 -3 -2 -1 -1 -2  0 -2
```

**Step 2: Assign Signs**

- $d_i > 0$: Positive sign (+)
- $d_i < 0$: Negative sign (−)
- $d_i = 0$: Discard the observation

Let $n$ be the number of non-zero differences.

```
# Remove zero differences

diff_nonzero <- diff[diff != 0]

# Assign signs

signs <- ifelse(diff_nonzero > 0, "+", "-")
data.frame(Difference = diff_nonzero, Sign = signs)
```

```
  Difference Sign
1         -2    -
2         -1    -
3          1    +
4         -3    -
5         -2    -
6         -1    -
7         -1    -
8         -2    -
9         -2    -
```

**Step 3: Count Signs**

- Number of positive signs: $n_+$
- Number of negative signs: $n_-$

```
n_pos <- sum(diff_nonzero > 0)
n_neg <- sum(diff_nonzero < 0)

n_pos
```

```
[1] 1
```

```
n_neg
```

```
[1] 8
```

Under $H_0$, positive and negative signs are equally likely.

**Step 4: Test Statistic**

The test statistic is defined as:

$$X = \min(n_+, n_-)$$

```
# Test statistic

X <- min(n_pos, n_neg)
X
```

```
[1] 1
```

Under the null hypothesis:

$$X \sim \text{Binomial}(n, 0.5)$$

```
# Perform Sign Test using binom.test()
binom.test(n_pos, n_pos + n_neg, p = 0.5, alternative = "two.sided")
```

```

    Exact binomial test

data:  n_pos and n_pos + n_neg
number of successes = 1, number of trials = 9, p-value = 0.03906
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.002809137 0.482496515
sample estimates:
probability of success
             0.1111111
```

**Step 5: Decision Rule**

- Compute the *p-value* using the binomial distribution
- Reject $H_0$ if:

$$\text{p-value} < \alpha$$

**Interpretation:**

- **Reject $H_0$:** There is sufficient evidence that the median difference is not zero, indicating that the calibration has a significant effect on defect counts.
- **Fail to reject $H_0$:** There is insufficient evidence to conclude that the calibration affects product defects.

The **Sign Test** is a reliable and robust nonparametric method for analyzing paired data when distributional assumptions are violated. Although it has lower power than alternative tests, it remains valuable in real-world applications where data quality is limited.

### 10.4.2 Wilcoxon Signed-Rank Test

The **Wilcoxon Signed-Rank Test** is a nonparametric test used to compare **paired samples** and test hypotheses about the **median of paired differences** [62]. Unlike the Sign Test, it considers both the **direction and magnitude** of differences, making it more powerful while still avoiding strict distributional assumptions.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/NZsL2eDQiDQ?si=r95mg9cdFBwlVVl5
```

The Wilcoxon Signed-Rank Test is commonly regarded as the **nonparametric alternative to the paired *t*-test**.

**The Wilcoxon Signed-Rank Test is appropriate when:**

- Observations are paired (before–after or matched samples)
- The distribution of differences is not normal
- Data are at least ordinal
- The distribution of differences is approximately symmetric
- The magnitude of changes is meaningful and reliable

**Key Characteristics:**

- Uses both the sign and rank of paired differences
- More powerful than the Sign Test
- Does not assume normality
- Sensitive to strong asymmetry

**Hypotheses**

$$H_0 : \text{Median difference} = 0$$

$$H_1 : \text{Median difference} \neq 0$$

**Real-World Case: Manufacturing Quality Control**

A manufacturing plant evaluates the effectiveness of a new **machine calibration procedure** designed to reduce product defects. For each production batch, the number of defective items is recorded **before** and **after** calibration.

Although the defect data are not normally distributed, the **magnitude of change** in defect counts is reliable and meaningful. Therefore, the **Wilcoxon Signed-Rank Test** is applied to assess whether the **median change in defect counts** differs significantly from zero.

**Step 1: Compute Paired Differences**

For each batch:

$$d_i = \text{Defects}_{\text{after}} - \text{Defects}_{\text{before}}$$

```
# Defect counts before and after calibration
before <- c(12, 15, 10, 18, 20, 14, 16, 22, 11, 19)
after  <- c(10, 14, 11, 15, 18, 13, 15, 20, 11, 17)

# Compute paired differences
diff <- after - before
diff
```

```
 [1] -2 -1  1 -3 -2 -1 -1 -2  0 -2
```

**Step 2: Remove Zero Differences**

- If $d_i = 0$, discard the observation
- Let $n$ be the number of non-zero differences

```
# Remove zero differences

diff_nonzero <- diff[diff != 0]
diff_nonzero
```

```
[1] -2 -1  1 -3 -2 -1 -1 -2 -2
```

**Step 3: Rank the Absolute Differences**

- Compute $|d_i|$ for each remaining pair
- Rank the values from smallest to largest
- If ties occur, assign average ranks

```
# Absolute differences

abs_diff <- abs(diff_nonzero)

# Rank absolute differences

ranks <- rank(abs_diff)

data.frame(
```

```
Difference = diff_nonzero,
AbsDifference = abs_diff,
Rank = ranks
)
```

```
  Difference AbsDifference Rank
1         -2             2  6.5
2         -1             1  2.5
3          1             1  2.5
4         -3             3  9.0
5         -2             2  6.5
6         -1             1  2.5
7         -1             1  2.5
8         -2             2  6.5
9         -2             2  6.5
```

**Step 4: Assign Signs to Ranks**

- If $d_i > 0$, assign a positive rank
- If $d_i < 0$, assign a negative rank

```
# Assign signed ranks

signed_ranks <- ifelse(diff_nonzero > 0, ranks, -ranks)

data.frame(
Difference = diff_nonzero,
Rank = ranks,
SignedRank = signed_ranks
)
```

```
  Difference Rank SignedRank
1         -2  6.5       -6.5
2         -1  2.5       -2.5
3          1  2.5        2.5
4         -3  9.0       -9.0
5         -2  6.5       -6.5
6         -1  2.5       -2.5
7         -1  2.5       -2.5
8         -2  6.5       -6.5
9         -2  6.5       -6.5
```

**Step 5: Compute Test Statistic**

Let:

- $W^+$ = sum of positive ranks
- $W^-$ = sum of negative ranks

The test statistic is defined as:

$$W = \min(W^+, W^-)$$

```
# Sum of positive and negative ranks

W_pos <- sum(ranks[diff_nonzero > 0])
W_neg <- sum(ranks[diff_nonzero < 0])

W_pos
```

```
[1] 2.5
```

```
W_neg
```

```
[1] 42.5
```

```
# Test statistic

W <- min(W_pos, W_neg)
W
```

```
[1] 2.5
```

For large samples ($n > 20$), the statistic may be approximated by a normal distribution.

**Step 6: Decision Rule**

Compute the *p-value* using the Wilcoxon distribution or its normal approximation

```
# Perform Wilcoxon Signed-Rank Test
wilcox.test(
after,
before,
paired = TRUE,
alternative = "two.sided",
exact = TRUE
)
```

```
    Wilcoxon signed rank test with continuity correction

data:  after and before
V = 2.5, p-value = 0.01868
alternative hypothesis: true location shift is not equal to 0
```

Reject $H_0$ if:

$$\text{p-value} < \alpha$$

**Interpretation:**

- **Reject $H_0$:** There is sufficient evidence that the median difference is not zero, indicating that the calibration significantly affects defect counts.
- **Fail to reject $H_0$:** There is insufficient evidence to conclude that the calibration has a significant effect.

The **Wilcoxon Signed-Rank Test** provides a balance between **robustness and efficiency**, making it a preferred nonparametric method for paired data when normality assumptions are violated.

### 10.4.3 Mann–Whitney U Test

The **Mann–Whitney U Test** is a nonparametric test used to compare **two independent samples** and assess whether they come from populations with the same **central tendency**. It is commonly regarded as the **nonparametric alternative to the independent two-sample *t*-test**.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/LcxB56PzylA?si=7ISh_4144R-9TF4P
```

Rather than comparing means, the Mann–Whitney U Test evaluates whether observations from one group tend to be **larger or smaller** than those from the other group based on their ranks.

**The Mann–Whitney U Test is appropriate when:**

- Two samples are **independent**
- Data are at least **ordinal**
- The population distributions are **not normal**
- Sample sizes may be small or unequal
- The shapes of the two distributions are similar

**Key Characteristics:**

- Uses ranks instead of raw data
- Does not assume normality
- Robust to outliers
- Tests differences in **distribution location**

**Hypotheses**

$$H_0 : \text{The two populations have the same distribution}$$

$$H_1 : \text{The two populations have different distributions}$$

(If distribution shapes are similar, this is often interpreted as a test of **median equality**.)

**Real-World Case: Business and Marketing Analytics**

A company wants to compare **customer satisfaction scores** between **two independent marketing strategies** (Strategy A and Strategy B). Survey responses are collected using a **Likert scale**, producing ordinal data that do not satisfy normality assumptions.

Because the two customer groups are independent and the data are ordinal, the **Mann–Whitney U Test** is used to determine whether customer satisfaction differs significantly between the two strategies.

**Step 1: Combine and Rank the Data**

- Combine observations from both groups
- Rank all observations from smallest to largest
- Assign average ranks in the presence of ties

```r
# Customer satisfaction scores (Likert scale: 1-5)
strategy_A <- c(3, 4, 4, 5, 3, 4, 5, 4)
strategy_B <- c(2, 3, 3, 4, 2, 3, 4, 3)

# Combine data
scores <- c(strategy_A, strategy_B)
group <- factor(c(rep("A", length(strategy_A)),
                  rep("B", length(strategy_B))))

# Rank combined data
ranks <- rank(scores)

data.frame(
  Score = scores,
  Group = group,
  Rank = ranks
)
```

```
   Score Group Rank
1      3     A  5.5
2      4     A 11.5
3      4     A 11.5
```

```
4       5     A 15.5
5       3     A  5.5
6       4     A 11.5
7       5     A 15.5
8       4     A 11.5
9       2     B  1.5
10      3     B  5.5
11      3     B  5.5
12      4     B 11.5
13      2     B  1.5
14      3     B  5.5
15      4     B 11.5
16      3     B  5.5
```

**Step 2: Compute Rank Sums**

Let:

- $R_1$ = sum of ranks for Group 1
- $R_2$ = sum of ranks for Group 2

```
# Rank sums

R1 <- sum(ranks[group == "A"])
R2 <- sum(ranks[group == "B"])

R1
```

```
[1] 88
```

```
R2
```

```
[1] 48
```

**Step 3: Compute the U Statistics**

For sample sizes $n_1$ and $n_2$:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

```r
# Sample sizes

n1 <- length(strategy_A)
n2 <- length(strategy_B)

# Compute U statistics

U1 <- n1 * n2 + n1 * (n1 + 1) / 2 - R1
U2 <- n1 * n2 + n2 * (n2 + 1) / 2 - R2

U1
```

```
[1] 12
```

```r
U2
```

```
[1] 52
```

**Step 4: Test Statistic**

The test statistic is:

$$U = \min(U_1, U_2)$$

```r
# Test statistic

U <- min(U1, U2)
U
```

```
[1] 12
```

For large samples, $U$ can be approximated by a **normal distribution**.

**Step 5: Decision Rule**

Compute the *p-value* from the Mann–Whitney distribution or its normal approximation

```r
# Mann-Whitney U Test using R
wilcox.test(
strategy_A,
strategy_B,
alternative = "two.sided",
exact = TRUE
)
```

```
    Wilcoxon rank sum test with continuity correction

data:  strategy_A and strategy_B
W = 52, p-value = 0.03033
alternative hypothesis: true location shift is not equal to 0
```

Reject $H_0$ if:

$$\text{p-value} < \alpha$$

**Interpretation:**

- **Reject $H_0$:** There is sufficient evidence that the two groups differ in central tendency or distribution location.
- **Fail to reject $H_0$:** There is insufficient evidence to conclude a difference between the two groups.

The **Mann–Whitney U Test** is a powerful and flexible nonparametric method for comparing two independent groups when parametric assumptions are violated or data are ordinal.

### 10.4.4 Kruskal–Wallis Test

The **Kruskal–Wallis Test** is a nonparametric test used to compare **three or more independent groups** and determine whether they originate from the same population distribution. It is commonly regarded as the **nonparametric alternative to one-way ANOVA**.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/l86wEhUzkY4?si=Vhkxk30XPJ1iSRdb
```

Instead of comparing group means, the Kruskal–Wallis Test evaluates differences in **central tendency** by comparing the **ranks** of observations across groups.

**The Kruskal–Wallis Test is appropriate when:**

- There are **three or more independent samples**
- Data are at least **ordinal**
- The population distributions are **not normal**
- Sample sizes may be unequal
- The shapes of the group distributions are similar

**Key Characteristics:**

- Uses ranks instead of raw data
- Does not assume normality

- Robust to outliers
- Suitable for small sample sizes

**Hypotheses**

$$H_0 : \text{All populations have the same distribution}$$

$$H_1 : \text{At least one population has a different distribution}$$

(If distribution shapes are similar, this is often interpreted as a test of **median equality**.)

**Real-World Case: Engineering and Quality Control**

A manufacturing company wants to compare **product defect rates** across **three different production machines** (Machine A, B, and C). The defect counts are skewed and contain outliers due to occasional machine malfunctions. Because the data are non-normal and the machines operate independently, the **Kruskal–Wallis Test** is applied to determine whether there are statistically significant differences in defect rates among the machines.

**Step 1: Combine and Rank All Observations**

- Combine observations from all groups into a single dataset
- Rank all values from smallest to largest
- Assign average ranks in the presence of ties

```
# Defect counts for each machine
machine_A <- c(5, 7, 6, 8, 9, 6, 7)
machine_B <- c(10, 12, 11, 9, 13, 10, 14)
machine_C <- c(4, 5, 6, 4, 5, 7, 6)

# Combine data
defects <- c(machine_A, machine_B, machine_C)
machine <- factor(c(rep("A", length(machine_A)),
                    rep("B", length(machine_B)),
                    rep("C", length(machine_C))))

# Rank all observations
ranks <- rank(defects)

data.frame(
  Defects = defects,
  Machine = machine,
  Rank = ranks
)
```

```
   Defects Machine Rank
1        5       A  4.0
2        7       A 11.0
3        6       A  7.5
4        8       A 13.0
5        9       A 14.5
6        6       A  7.5
7        7       A 11.0
8       10       B 16.5
9       12       B 19.0
10      11       B 18.0
11       9       B 14.5
12      13       B 20.0
13      10       B 16.5
14      14       B 21.0
15       4       C  1.5
16       5       C  4.0
17       6       C  7.5
18       4       C  1.5
19       5       C  4.0
20       7       C 11.0
21       6       C  7.5
```

**Step 2: Compute Rank Sums for Each Group**

Let:

- $R_j$ = sum of ranks for group $j$
- $n_j$ = sample size of group $j$
- $k$ = number of groups
- $N = \sum_{j=1}^{k} n_j$

```
# Sample sizes
n_A <- length(machine_A)
n_B <- length(machine_B)
n_C <- length(machine_C)

# Rank sums
R_A <- sum(ranks[machine == "A"])
R_B <- sum(ranks[machine == "B"])
R_C <- sum(ranks[machine == "C"])

R_A
```

```
[1] 68.5
```

```
R_B
```

```
[1] 125.5
```

```
R_C
```

```
[1] 37
```

**Step 3: Compute the Test Statistic**

The Kruskal–Wallis test statistic is:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

```
# Total sample size
N <- length(defects)

# Compute H statistic
H <- (12 / (N * (N + 1))) * (
(R_A^2 / n_A) +
(R_B^2 / n_B) +
(R_C^2 / n_C)
) - 3 * (N + 1)

H
```

```
[1] 14.93321
```

**Step 4: Sampling Distribution**

For sufficiently large samples, the test statistic follows a **chi-square distribution**:

$$H \sim \chi^2_{k-1}$$

```
# Degrees of freedom
df <- 3 - 1

# Compute p-value
p_value <- 1 - pchisq(H, df)
p_value
```

```
[1] 0.0005718666
```

**Step 5: Decision Rule**

Compute the *p-value* from the chi-square distribution

```
# Kruskal-Wallis test using R
kruskal.test(defects ~ machine)
```

```
    Kruskal-Wallis rank sum test

data:  defects by machine
Kruskal-Wallis chi-squared = 15.14, df = 2, p-value = 0.0005158
```

Reject $H_0$ if:

$$\text{p-value} < \alpha$$

**Interpretation:**

- **Reject $H_0$:** There is sufficient evidence that at least one group differs in central tendency or distribution location.
- **Fail to reject $H_0$:** There is insufficient evidence to conclude a difference among the groups.

**Post-Hoc Analysis**

If $H_0$ is rejected, post-hoc tests such as **Dunn's test** may be conducted to identify which specific groups differ. The **Kruskal–Wallis Test** provides a robust and flexible approach for comparing multiple independent groups when parametric assumptions are violated.

### 10.4.5 Friedman Test

The **Friedman Test** is a nonparametric statistical test used to detect differences among **three or more related (paired) groups**. It is commonly regarded as the **nonparametric alternative to the one-way repeated-measures ANOVA**.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/2moNzzkkZwU?si=au5hcqwTJyb0zsQf
```

Rather than comparing means, the Friedman Test evaluates differences in **central tendency** by comparing **within-subject ranks** across treatments or conditions.

**The Friedman Test is appropriate when:**

- The same subjects are measured under **three or more conditions**
- Observations are **paired or repeated measures**
- Data are at least **ordinal**

- The normality assumption for repeated-measures ANOVA is violated
- The magnitude of measurements is comparable across conditions

**Key Characteristics:**

- Uses ranks within each subject/block
- Does not assume normality
- Controls for subject-to-subject variability
- Suitable for small sample sizes

**Hypotheses**

$$H_0 : \text{All treatments have the same distribution}$$

$$H_1 : \text{At least one treatment has a different distribution}$$

(If distribution shapes are similar, this is often interpreted as a test of **median equality** across treatments.)

**Real-World Case: Human Performance Evaluation**

A company evaluates **employee productivity** under **three different work schedules**: fixed hours, flexible hours, and remote work. The **same employees** are evaluated under each schedule over separate periods.

Because the productivity scores are not normally distributed and measurements are repeated on the same individuals, the **Friedman Test** is used to determine whether productivity differs significantly across the three work conditions.

**Step 1: Organize Data into Blocks**

- Each **row** represents a subject (block)
- Each **column** represents a treatment or condition

```
# Defect counts for each batch under different settings
setting_A <- c(8, 7, 9, 6, 10)
setting_B <- c(6, 5, 7, 5, 8)
setting_C <- c(9, 8, 10, 7, 11)

# Create data frame (blocks = batches)
defects <- data.frame(
  Batch = factor(1:5),
  A = setting_A,
  B = setting_B,
```

```
  C = setting_C
)

defects
```

```
  Batch  A B  C
1     1  8 6  9
2     2  7 5  8
3     3  9 7 10
4     4  6 5  7
5     5 10 8 11
```

**Step 2: Rank Data Within Each Block**

- Rank the values **within each subject** from smallest to largest
- Assign average ranks in case of ties

```
# Rank within each batch
ranks <- t(apply(defects[, -1], 1, rank))

colnames(ranks) <- c("A", "B", "C")
ranks
```

```
     A B C
[1,] 2 1 3
[2,] 2 1 3
[3,] 2 1 3
[4,] 2 1 3
[5,] 2 1 3
```

**Step 3: Compute Rank Sums for Each Treatment**

Let:

- $R_j$ = sum of ranks for treatment $j$
- $n$ = number of subjects (blocks)
- $k$ = number of treatments

```
# Rank sums
R <- colSums(ranks)

n <- nrow(defects)   # number of blocks
k <- ncol(ranks)     # number of treatments

R
```

```
 A  B  C
10  5 15
```

**Step 4: Compute the Test Statistic**

The Friedman test statistic is:

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

```
# Compute Friedman statistic manually
Q <- (12 / (n * k * (k + 1))) * sum(R^2) - 3 * n * (k + 1)
Q
```

```
[1] 10
```

**Step 5: Sampling Distribution**

For sufficiently large samples, the test statistic follows a **chi-square distribution**:

$$Q \sim \chi^2_{k-1}$$

```
# Degrees of freedom
df <- k - 1

# Compute p-value
p_value <- 1 - pchisq(Q, df)
p_value
```

```
[1] 0.006737947
```

**Step 6: Decision Rule**

Compute the *p-value* from the chi-square distribution

```
# Friedman test using R
friedman.test(as.matrix(defects[, -1]))
```

```
    Friedman rank sum test

data:  as.matrix(defects[, -1])
Friedman chi-squared = 10, df = 2, p-value = 0.006738
```

Reject $H_0$ if:

$$\text{p-value} < \alpha$$

**Interpretation:**

- **Reject $H_0$:** There is sufficient evidence that at least one treatment differs in central tendency.
- **Fail to reject $H_0$:** There is insufficient evidence to conclude a difference among treatments.

**Post-Hoc Analysis**

If $H_0$ is rejected, post-hoc procedures such as the **Nemenyi test** or pairwise **Wilcoxon signed-rank tests** with adjustment may be used to identify which treatments differ.

The **Friedman Test** is a powerful nonparametric method for analyzing repeated-measures or blocked data when parametric assumptions are violated.

### 10.4.6 Chi-Square Test

The **Chi-Square Test** is a nonparametric statistical test used to examine relationships between **categorical variables**. It is widely applied to determine whether observed frequencies differ significantly from expected frequencies under a specified hypothesis.

```
Video cannot be displayed in PDF/Word.
 Please view the HTML version or open directly on YouTube:
 https://www.youtube.com/embed/EjtXk-yEK6w?si=qjzHP61u-HmxRras
```

The Chi-Square Test is commonly used for **independence testing** and **goodness-of-fit analysis**.

**The Chi-Square Test is appropriate when:**

- Data are **categorical**
- Observations are **independent**
- Frequencies (counts) are analyzed, not percentages
- Expected frequencies in each cell are sufficiently large (typically $\geq 5$)

**Key Characteristics:**

- Based on frequency counts
- Does not assume normality
- Simple to compute and interpret
- Sensitive to sample size

**Types of Chi-Square Tests**

1. **Chi-Square Test of Independence**
   Examines whether two categorical variables are associated.
2. **Chi-Square Goodness-of-Fit Test**
   Determines whether an observed distribution matches a theoretical distribution.

**Hypotheses (Independence Test)**

$$H_0 : \text{The two categorical variables are independent}$$

$$H_1 : \text{The two categorical variables are not independent}$$

**Real-World Case: Social and Behavioral Sciences**

A university wants to examine whether **students' study programs** (Science, Engineering, Social Sciences) are associated with their **preferred learning mode** (online, hybrid, in-person).

Because both variables are categorical and the data consist of frequency counts, the **Chi-Square Test of Independence** is applied.

**Step 1: Construct a Contingency Table**

| Program / Learning Mode | Online | Hybrid | In-Person |
|---|---|---|---|
| Science | $O_{11}$ | $O_{12}$ | $O_{13}$ |
| Engineering | $O_{21}$ | $O_{22}$ | $O_{23}$ |
| Social Sciences | $O_{31}$ | $O_{32}$ | $O_{33}$ |

```r
# Observed frequencies
observed <- matrix(
  c(40, 35, 25,   # Science
    30, 45, 25,   # Engineering
    50, 30, 20),  # Social Sciences
  nrow = 3,
  byrow = TRUE
)

colnames(observed) <- c("Online", "Hybrid", "In-Person")
rownames(observed) <- c("Science", "Engineering", "Social Sciences")

observed
```

```
                Online Hybrid In-Person
Science             40     35        25
Engineering         30     45        25
Social Sciences     50     30        20
```

**Step 2: Compute Expected Frequencies**

For each cell:

$$E_{ij} = \frac{(\text{Row Total})_i \times (\text{Column Total})_j}{\text{Grand Total}}$$

```
# Compute expected frequencies
expected <- chisq.test(observed)$expected
expected
```

```
                Online   Hybrid In-Person
Science             40 36.66667  23.33333
Engineering         40 36.66667  23.33333
Social Sciences     40 36.66667  23.33333
```

**Step 3: Compute the Test Statistic**

The Chi-Square statistic is:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- $O_{ij}$ = observed frequency
- $E_{ij}$ = expected frequency
- $r$ = number of rows
- $c$ = number of columns

```
# Chi-square statistic
chisq_stat <- sum((observed - expected)^2 / expected)
chisq_stat
```

```
[1] 8.896104
```

**Step 4: Degrees of Freedom**

$$df = (r-1)(c-1)$$

```
# Degrees of freedom
df <- (nrow(observed) - 1) * (ncol(observed) - 1)
df
```

```
[1] 4
```

**Step 5: Decision Rule**

Compute the *p-value* from the chi-square distribution

```
# Chi-Square Test of Independence using R
chisq.test(observed)
```

```
    Pearson's Chi-squared test

data:  observed
X-squared = 8.8961, df = 4, p-value = 0.06375
```

Reject $H_0$ if:

$$\text{p-value} < \alpha$$

**Interpretation**

- **Reject $H_0$:** There is a significant association between the categorical variables.
- **Fail to reject $H_0$:** There is insufficient evidence to conclude an association.

**Effect Size (Optional)**

For contingency tables, effect size can be measured using **Cramér's V**:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where $k$ is the smaller of $r$ or $c$.

The **Chi-Square Test** is a fundamental tool for analyzing categorical data and identifying relationships between qualitative variables in many applied research fields.

## 10.5 Advantages and Limitations

The following table presents a summary of the advantages and limitations of **nonparametric statistical methods**, which may be considered when selecting an appropriate analytical approach based on data characteristics and research objectives.

| Method | Advantages | Disadvantages |
|---|---|---|
| Sign Test | Fewer assumptions; extremely robust to outliers | Very low statistical power; ignores magnitude of differences |
| Wilcoxon Signed-Rank Test | More powerful than Sign Test; uses magnitude and direction | Requires symmetric distribution of differences |

| Method | Advantages | Disadvantages |
|---|---|---|
| Mann–Whitney U Test | Suitable for ordinal data; robust to non-normality | Tests distributional shift rather than mean difference |
| Kruskal–Wallis Test | Extends Mann–Whitney to multiple groups; no normality assumption | Does not indicate which groups differ without post-hoc tests |
| Friedman Test | Suitable for repeated measures; controls subject variability | Less powerful than parametric repeated-measures ANOVA |
| Chi-Square Test | Ideal for categorical data; simple and intuitive | Sensitive to small expected frequencies; no direction of association |

## 10.6 Nonparametric Case Studies

This section presents several **real-world case studies** illustrating the application of **nonparametric statistical methods** in different fields. Each case highlights the characteristics of the data, the rationale for choosing a nonparametric approach, and the appropriate statistical test used to address the research question.

### 10.6.1 Case Study 1

**Manufacturing Quality Control (Sign Test):**

A manufacturing plant investigates whether a new **machine calibration procedure** reduces the number of defective products. Defect counts are recorded **before and after** calibration for the same production batches. The data contain extreme values due to occasional machine failures and do not satisfy normality assumptions. **Objective:** Test whether the median change in defect counts differs from zero.

### 10.6.2 Case Study 2

**Medical Treatment Evaluation (Wilcoxon Signed-Rank Test):**

A clinical researcher examines whether a new therapy reduces **patient pain scores** measured on an ordinal scale. Pain levels are recorded for the same patients **before and after** treatment. The distribution of differences is non-normal, but the magnitude of change is meaningful. **Objective:** Determine whether the median pain score after treatment differs from before treatment.

### 10.6.3 Case Study 3

**Marketing Strategy Comparison (Mann–Whitney U Test):**

A company compares **customer satisfaction ratings** between two independent marketing strategies. Survey responses are collected using a Likert scale from two separate customer groups. **Objective:** Assess whether customer satisfaction differs between the two strategies.

### 10.6.4 Case Study 4

**Production Line Performance (Kruskal–Wallis Test):**

An engineering team evaluates **defect rates** across **three independent production machines**. The defect data are skewed and contain outliers. **Objective:** Identify whether at least one machine has a different defect rate distribution.

### 10.6.5 Case Study 5

**Human Performance Analysis (Friedman Test):**

An organization studies employee productivity under **three different work conditions** (onsite, hybrid, remote). Productivity scores are measured for the **same employees** under each condition. **Objective:** Determine whether productivity differs across work conditions.

### 10.6.6 Case Study 6

**Education and Learning Preferences (Chi-Square Test):**

A university analyzes the relationship between **students' study programs** and their **preferred learning modes** (online, hybrid, in-person). Data are collected as frequency counts. **Objective:** Examine whether learning preferences are associated with study programs.

These case studies demonstrate how nonparametric methods provide flexible and robust solutions when data violate parametric assumptions or involve ordinal and categorical measurements.

## References

[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., Introduction to the practice of statistics, Macmillan Learning, 2020

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

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

[8] Baker, L., Data types: Getting started with statistics, Everand, 2020, Available. https://www.everand.com/book/486797108/Data-Types-Getting-Started-With-Statistics

[9] Shreffler, J., Types of variables and commonly used statistical designs, *NCBI Bookshelf*, 2023, Available. https://www.ncbi.nlm.nih.gov/books/NBK557882/

[10] Baley, I. and Veldkamp, L., The data economy: Tools and applications, Princeton University Press, 2025, Available. https://press.princeton.edu/books/hardcover/9780691256726/the-data-economy

[11] MyGreatLearning, 4 types of data - nominal, ordinal, discrete, continuous, Available. https://www.mygreatlearning.com/blog/types-of-data/

[12] Canada, S., 4.2 types of variables, 2021, Available. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch8/5214817-eng.htm

[13] Adelaide, U. of, Types of data in statistics: Numerical vs categorical data, Available. https://online.adelaide.edu.au/blog/types-of-data

[14] Scribbr, Types of variables in research & statistics | examples, 2022, Available. https://www.scribbr.com/methodology/types-of-variables/

[15] GeeksforGeeks, Data types in statistics, 2025, Available. https://www.geeksforgeeks.org/maths/data-types-in-statistics/

[16] James, G., Witten, D., Hastie, T., and Tibshirani, R., An introduction to statistical learning: With applications in r, Springer, 2021

[17] Wakeling, I., Statistics in r using RStudio: An introduction for food scientists, Elsevier, Cambridge, MA, 2020

[18] Hastie, T., Tibshirani, R., and Friedman, J., The elements of statistical learning: Data mining, inference, and prediction, Springer, 2021

[19] Wickham, H., Tidy data: A practical guide to organizing and managing data in r, O'Reilly Media, Sebastopol, CA, 2023

[20] Kelleher, J. D. and Tierney, B., Data science: An introduction, CRC Press, Boca Raton, FL, 2015

[21] Cairo, A., How charts lie: Getting smarter about visual information, W. W. Norton & Company, New York, 2023

[22] Healy, K., Data visualization: A practical introduction, Princeton University Press, Princeton, NJ, 2022

[23] Wilke, C. O., Fundamentals of data visualization: A primer on making informative and compelling figures, O'Reilly Media, Sebastopol, CA, 2019

[24] Wilke, C. O., Fundamentals of data visualization, O'Reilly Media, Sebastopol, CA, 2019

[25] Tufte, E. R., The visual display of quantitative information, Graphics Press, Cheshire, CT, 2001

[26] Knaflic, C. N., Storytelling with data: A data visualization guide for business professionals, Wiley, Hoboken, NJ, 2015

[27] Wickham, H., ggplot2: Elegant graphics for data analysis, Springer-Verlag New York, New York, NY, 2016

[28] Alooba, M., Visual analytics for business intelligence: Techniques and applications, *Journal of Business Analytics*, vol. 10, no. 2, 45–60, 2023

[29] BoldBI, Boxplots and scatter plots for outlier detection, https://www.boldbi.com/learning-center/data-visualization/boxplot-scatterplot, 2023

[30] Domo, Understanding data symmetry, skewness, and distribution, https://www.domo.com/learn/data-visualization, 2023

[31] Yi, L., Bar charts: Understanding and visualizing categorical data, https://towardsdatascience.com, 2023

[32] WebDataRocks, Bar chart explained: Definition, examples, and use cases, https://www.webdatarocks.com, 2022

[33] Codecademy, How to read and make bar charts, https://www.codecademy.com, 2023

[34] Atlassian, Data visualization basics: Histograms and bar charts, https://www.atlassian.com, 2023

[35] GeeksforGeeks, Understanding histograms in data visualization, https://www.geeksforgeeks.org, 2025

[36] JMP, Using histograms to explore data distributions, https://www.jmp.com, 2023
[37] Tableau, Histogram: Definition and use cases in data visualization, https://www.tableau.com/learn/articles/histogram, 2023
[38] Tableau, When to use pie charts and when not to, https://www.tableau.com/learn/articles/pie-chart, 2023
[39] Datawrapper Academy, How to create better pie charts, https://blog.datawrapper.de/pie-charts/, 2024
[40] Statistics How To, Statistics how to: Simple definitions and formulas, 2024, Available. https://www.statisticshowto.com/
[41] OpenStax, Introductory statistics, OpenStax, Rice University, 2023, Available. https://openstax.org/books/introductory-statistics/pages/1-introduction
[42] Jamovi Project, Jamovi statistical software guide, 2024, Available. https://www.jamovi.org
[43] Mann, P. S., Introductory statistics, John Wiley & Sons, Hoboken, NJ, 2010
[44] Triola, M. F., Elementary statistics, Pearson, Boston, MA, 2018
[45] Field, A., Discovering statistics using IBM SPSS statistics, SAGE Publications, London, 2013
[46] Moore, D. S., McCabe, G. P., and Craig, B. A., Introduction to the practice of statistics, W. H. Freeman; Company, New York, 2018
[47] Freedman, D., Pisani, R., and Purves, R., Statistics, W. W. Norton & Company, New York, 2007
[48] Jim Frost, Understanding variability in statistics, 2023, Available. https://statisticsbyjim.com/basics/variability/
[49] Notes, P. H., Understanding measures of dispersion in public health statistics, 2020, Available. https://www.publichealthnotes.com/measures-of-dispersion/
[50] Columbia University Mailman School of Public Health, Measures of dispersion, 2021, Available. https://www.publichealth.columbia.edu/research/population-health-methods/measures-dispersion
[51] GeeksforGeeks, Mean, variance, and standard deviation in statistics, 2022, Available. https://www.geeksforgeeks.org/mean-variance-and-standard-deviation-in-statistics/
[52] Diez, D. M., Barr, C. D., and Çetinkaya-Rundel, M., OpenIntro statistics, OpenIntro, Boston, MA, 2019, Available. https://www.openintro.org/book/os/
[53] Moore, D. S., Notz, W. I., and Fligner, M. A., The basic practice of statistics, W. H. Freeman; Company, New York, NY, 2017
[54] Triola, M. F., Elementary statistics, Pearson Education, Boston, MA, 2018
[55] Gravetter, F. J., Wallnau, L. B., and Forzano, L.-A. B., Essentials of statistics for the behavioral sciences, Cengage Learning, Boston, MA, 2021
[56] Hogg, R. V., Tanis, E. A., and Zimmerman, D. L., Probability and statistical inference, global edition, Pearson Education, London, 2024
[57] Ghahramani, S., Fundamentals of probability, CRC Press, Boca Raton, 2024
[58] Barron, E. N. and Del Greco, J. G., Probability and statistics for STEM: A course in one semester, Springer, Cham, 2024
[59] Jim Frost, Statistical inference overview, 2023, Available. https://statisticsbyjim.com/hypothesis-testing/statistical-inference/
[60] Khan Academy, Statistical inference, 2023, Available. https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample
[61] GeeksforGeeks, Statistical inference in statistics, 2023, Available. https://www.geeksforgeeks.org/statistical-inference/

[62] Conover, W. J., Practical nonparametric statistics, John Wiley & Sons, New York, 1999
[63] Gibbons, J. D. and Chakraborti, S., Nonparametric statistical inference, Chapman; Hall/CRC, Boca Raton, 2011
[64] Hollander, M., Wolfe, D. A., and Chicken, E., Nonparametric statistical methods, John Wiley & Sons, Hoboken, 2014