

my-ebook

Parmeshvar

2025-07-06

Table of contents

1	Introduction	8
1.1	Teaching	8
1.2	Lecture notes	9
1.3	Moodle website	9
2	Week 1	13
2.1	Module 1: Introduction to Statistics	13
2.1.1	Pre-Requisites	13
2.1.2	Agenda	13
2.1.3	Meaning of Statistics	13
2.1.4	Nature of Statistics	14
2.1.5	Uses of Statistics	14
2.1.6	Limitations of Statistics	14
2.2	Misuse of Statistics	14
2.3	Module 2: Mathematics vs Statistics	15
2.4	Module 3: Software-Based Statistical Revolution	15
2.4.1	Popular Statistical Software	15
2.4.2	GUI vs CLI	15
2.4.3	Recommended GUI Tools for R	16
2.4.4	Installing RKWard on Ubuntu	16
2.5	Module 4: Understanding Variables	16
2.5.1	What is a Variable?	16
2.5.2	R Definition:	16
3	Week 2	23
3.1	Introduction	23
3.1.1	Purpose of the eBook	23
3.1.2	Who Should Read This?	23
3.1.3	What You'll Learn	23
3.2	1. Fundamentals of Statistics	23
3.2.1	1.1 What is Statistics?	23
3.2.2	1.2 Key Objectives	24
3.2.3	1.3 Types of Statistics	24
3.3	2. Types of Data	24
3.3.1	2.1 Classification of Data	24
3.4	3. Descriptive Statistics	25
3.4.1	3.1 Measures of Central Tendency	25
3.4.2	3.2 The Mean	25
3.4.3	3.3 The Median	26
3.4.4	3.4 The Mode	26

3.4.5	3.5 Comparison Table	27
3.5	4. Measures of Variability	27
3.5.1	4.1 Why Measure Variability?	27
3.5.2	4.2 Range	27
3.5.3	4.3 Quartiles and Interquartile Range	28
3.5.4	4.4 Variance	28
3.5.5	4.5 Standard Deviation	29
3.5.6	4.6 Coefficient of Variation (CV)	29
3.5.7	4.7 Moment-Based Measures	29
3.6	5. Probability Fundamentals	30
3.6.1	5.1 Introduction to Probability	30
3.6.2	5.2 Key Definitions	30
3.6.3	5.3 Types of Events	30
3.6.4	5.4 Classical Probability	30
3.6.5	5.5 Probability Rules	30
3.6.6	5.6 Conditional Probability	31
3.7	6. Discrete Probability Distributions	32
3.7.1	6.1 Bernoulli Distribution	32
3.7.2	6.2 Binomial Distribution	32
3.8	7. Continuous Distributions	33
3.8.1	7.1 Normal Distribution	33
3.8.2	7.2 Standard Normal Distribution	33
3.9	8. Visualizing Data	34
3.9.1	8.1 Frequency Distribution	34
3.9.2	8.2 Histogram	34
3.9.3	8.3 Boxplot (Box-and-Whisker Plot)	34
3.9.4	8.4 Scatter Plot	34
3.10	9. Practical Applications	35
3.10.1	9.1 Business Use Cases	35
3.10.2	9.2 Education and Research	35
3.11	10. Using RKWard	35
3.11.1	10.1 What is RKWard?	35
3.11.2	10.2 Installation Guide	35
3.11.3	10.3 Sample RKWard Activities	35
3.11.4	10.4 Using R Code in RKWard	36
4	Week 3	37
5	Introduction	38
5.1	Importance of Statistics	38
5.2	Overview of Topics	38
6	Understanding Populations and Samples	39
6.1	Population	39
6.2	Sample	39
6.3	Why Use Samples?	39
6.4	Relation Between Population & Sample	39

7	Hypotheses and Errors	40
7.1	Hypothesis Defined	40
7.1.1	Null Hypothesis (H_0)	40
7.1.2	Alternative Hypothesis (H_A)	40
7.2	Types of Errors	40
7.3	Significance Level ()	40
8	Inferential Statistics	41
8.1	Purpose	41
8.2	Common Techniques	41
8.3	Sampling Techniques	41
8.3.1	1. Simple Random Sampling	41
8.3.2	2. Systematic Sampling	41
8.3.3	3. Stratified Sampling	41
8.3.4	4. Cluster Sampling	41
8.4	Central Limit Theorem (CLT)	42
9	Descriptive Statistics	43
9.1	Measures of Central Tendency	43
9.1.1	Mean	43
9.1.2	Median	43
9.1.3	Mode	43
9.2	Measures of Dispersion	43
9.2.1	Range	43
9.2.2	Variance	43
9.2.3	Standard Deviation	43
9.3	Measures of Shape	44
10	Graphical Methods	45
10.1	Histogram	45
10.2	R Data Types and Structures	45
10.3	Comparing R vs Excel vs GUI-R (RKWard)	45
10.4	Installing RKWard (Ubuntu)	46
10.5	Teaching Tools in RKWard	46
10.6	GUI-Based Statistical Tools	46
11	Linear Regression in R	47
11.1	What is Linear Regression?	47
11.1.1	Simple Linear Regression Equation:	47
11.2	Code Example	47
12	Fit model	48
13	Summary	49
13.1	Adjusted R-squared	49
13.2	Normal Distribution	49
13.3	Data Import Techniques	49
13.4	Working with the RKWard Interface	49
13.5	Spreadsheet Concepts	50

13.6 Advantages	50
13.7 Limitations	50
13.8 Advanced Plots and Techniques	50
13.9 Common R Packages for Statistics	52
13.10 Introduction to Command Line	52
13.11 Windows Terminal	52
13.12 Git + R Project Example	53
13.13 R Script Template for Analysis	53
14 Load data	54
15 Descriptive stats	55
16 Histogram	56
17 Linear regression	57
18 Scatter plot with regression line	58
18.1 Future Applications of Statistics	58
18.2 Practice Challenges	58
18.3 Key Takeaways	58
19 Week 4	59
20 Introduction	60
21 Course Overview	61
21.1 Course Name	61
21.2 Instructor Profile	61
21.3 Learning Objectives	61
22 Chapter 1: Fundamental Concepts	62
22.1 Descriptive Statistics	62
22.1.1 Central Tendency	62
22.1.2 Dispersion	62
22.2 Standard Error	63
22.3 Central Limit Theorem	63
22.4 Confidence Intervals	63
23 Chapter 2: Estimation	64
23.1 Types of Estimates	64
23.2 Parameter vs Statistic	64
24 Chapter 3: Hypothesis Testing	65
25 Chapter 4: Student's T-Test	66
25.1 Types	66
25.2 One-Sample T-Test Example	66
25.3 Test Statistic	66
25.4 Degrees of Freedom	66

25.5	Decision Rule	67
25.6	T-Test in GUI-R	67
26	Chapter 5: ANOVA	68
26.1	Purpose	68
26.1.1	One-Way ANOVA Formula	68
26.1.2	Assumptions	68
26.1.3	Example Table	68
26.2	Post-Hoc Tests	68
26.3	ANOVA in GUI-R	69
27	Chapter 6: GUI-R Workflow	70
28	Chapter 7: Advanced Concepts	71
28.1	Variance Partitioning	71
28.2	Degrees of Freedom	71
28.3	Chi-Square and F Distribution	71
28.4	Univariate, Bivariate, Multivariate	71
28.5	Parametric Test Assumptions	72
28.6	Effect Size	72
28.7	Power of a Test	72
29	Conclusion	73
30	References	74
31	Chapter 8: Advanced T-Test Applications	75
31.1	Paired Sample T-Test	75
31.1.1	Example:	75
31.2	Independent Samples T-Test	75
31.3	One-Sample T-Test with GUI-R	76
32	Chapter 9: More on Confidence Intervals	77
32.1	Visualizing Confidence Intervals in R	77
33	Chapter 10: Robust ANOVA Models	78
33.1	Two-Way ANOVA	78
33.2	Repeated Measures ANOVA	78
34	Chapter 11: Effect Size Measures	80
34.1	Cohen's d	80
34.1.1	R Example	80
34.2	Eta-Squared (η^2)	80
35	Chapter 12: Statistical Assumptions Checking	81
35.1	Normality	81
35.2	Homogeneity of Variance	82
36	Chapter 13: Non-Parametric Alternatives	83
36.1	Wilcoxon Signed Rank Test	83

36.2 Mann-Whitney U Test	83
36.3 Kruskal-Wallis Test	83
37 Chapter 14: Visualizing Statistical Results	85
37.1 Boxplots	85
37.2 Histograms	85
37.3 Density Plot	86
38 Chapter 15: RKWard (GUI-R) Tips	87
38.1 Summary: Basic Statistics using GUI-R (RKWard)	87

1 Introduction

Introduction

DR.Harsh Pradhan, Phone: +91-9930034241 , Email: harsh.231284@gmail.com, Institute of Management Studies, Banaras Hindu University, Address: 18-GF, Jaipuria Enclave, Kaushambhi, Ghaziabad, India, 201010

Interest: Goal Orientation Job Performance Consumer Behavior Behavioral Finance Bibliometric Analysis Options as Derivatives Statistics Indian Knowledge System,

[Orcid ID](#)

[Google Scholar](#)

[GitHub](#)

[Researcher ID](#)

[Personal Website](#)

[Youtube ID](#)

Doing a PhD with me: [README.1st](#)

[Academic Profile](#)

Topics to be covered:

1. Basic probability theory, random variable theory (including jointly distributed RVs), probability distributions (including bivariate distributions)
2. Using Bayes' rule for statistical inference
3. An introduction to (generalized) linear models
4. An introduction to hierarchical models
5. Measurement error models
6. Mixture models
7. Model selection and hypothesis testing (Bayes factor and k-fold cross-validation)

1.1 Teaching

Science and statistics is/are one unitary thing; you cannot do one without the other. Towards this end, I teach some (in my opinion) critically important classes that provide a solid statistical foundation for doing research in cognitive science.

Courses offered:

1. Free online course, four weeks (MOOC), enrollments open: Introduction to Bayesian Data Analysis
2. Short (four-hour) tutorial on Bayesian statistics, taught at EMLAR 2022: [here](#)
3. Introduction to (frequentist) statistics
4. Introduction to Bayesian data analysis for cognitive science
5. BDA cover

1.2 Lecture notes

Download from [here](#).

1.3 Moodle website

All communications with students in Potsdam will be done through [this website](#). # Schedule

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
Week 1	2	Descriptive Statistics	Central Tendency	Video	Week 2.pdf
	2	Descriptive Statistics	Measure of Variability	Video	Same as above
	3	Descriptive Statistics	Describing Data	Video	Same as above
	4	Descriptive Statistics	Probability	Video	Same as above
	5	Descriptive Statistics	Distribution	Video	Same as above
Week 3	1	Descriptive Statistics	Z Table (Normal Distribution)	Video	Week 3.pdf
	2	Descriptive Statistics	Measuring Divergence	Video	Same as above
	3	Inferential Statistics	Sample and Population	Video	Same as above
	4	Inferential Statistics	Model Fit	Video	Same as above

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
	5	Inferential Statistics	Hypothesis and Error	Video	Same as above
Week 4	1	Terms of Statistics	Terms of Statistics	Video	Week 4.pdf
	2	Terms of Statistics	T-Test	Video	Same as above
	3	Terms of Statistics	T-Test in Detail	Video	Same as above
	4	ANOVA	ANOVA	Video	Same as above
Week 5	1	ANOVA	Example of ANOVA	Video	Week 5.pdf
	2	ANOVA	Types of ANOVA	Video	Same as above
	3	Correlation	Introduction to Correlation	Video	Same as above
	4	Correlation	Regression (Part 1)	Video	Same as above
	5	Correlation	Regression (Part 2)	Video	Same as above
Week 6	1	Correlation	R Script for Regression	Video	Week 6.pdf
	2	Chi Square	Chi Square	Video	Same as above
	3	Chi Square	Chi Square Test	Video	Same as above
	4	Logistic Function	Regression Function	Video	Same as above
	5	Logistic Function	Distribution	Video	Same as above
Week 7	1	Time Series	Intro to Time Series	Video	Week 7.pdf
	2	Time Series	Conditional Probability	Video	Same as above
	3	Time Series	Additional Concepts	Video	Same as above
	4	Time Series	Distribution	Video	Same as above
	5	Time Series	Poisson Distribution	Video	Same as above

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
	6	Index Numbers	Price & Quantity Index	Video	Same as above
	7	Decision Environments	Risk/Uncertainty, Bayes, Trees	Video	Same as above
	8	Time Series Analysis	Components, Trend, Seasonality	Video	Same as above
	9	Time Series Analysis	Least Squares Method	Video	Same as above
Week 8	1	Effect Size & Documentation	Package/Library	Video	Week 8.pdf
	2	Effect Size & Documentation	RStudio vs RKward	Video	Same as above
	3	Effect Size & Documentation	Flexplot	Video	Same as above
	4	Effect Size & Documentation	Functions	Video	Same as above
	5	Effect Size & Documentation	R Shiny & R Markdown	Video	Same as above
	6	Effect Size & Documentation	Application with Real Datasets	Video	Same as above

Week	Main Topic	Subtopic	Video	PDF Resource
7	Effect Size & Interpretation	Importance in Testing	Video	Same as above
8	Effect Size & Interpretation	Installing dplyr, ggplot2	Video	Same as above
9	Effect Size & Interpretation	Visual Model Interpretation	Video	Same as above
10	Effect Size & Interpretation	Creating/Using Functions	Video	Same as above
11	Effect Size & Interpretation	Report, Dashboard, Interactivity	Video	Same as above

2 Week 1

2.1 Module 1: Introduction to Statistics

2.1.1 Pre-Requisites

- Just an open and eager mind
- Basic understanding of Mathematics or Statistics

2.1.2 Agenda

- Meaning of Statistics
 - Nature and Scope
 - Uses of Statistics
 - Limitations
 - Fallacies and Misuse
 - Math vs Statistics
 - GUI Tools & Transition to Software-based Stats
-

2.1.3 Meaning of Statistics

Statistics is a science which provides tools for **analysis and interpretation** of raw data collected for decision-making in diverse fields.

It includes four core concepts:

- **Population** – Complete data or total group
- **Sample** – Subset of population
- **Parameter** – Numerical summary from population
- **Statistic** – Numerical summary from sample

2.1.4 Nature of Statistics

- Deals with **numerical facts**
 - Focused on **social phenomena** and real-world data
 - Organizes, classifies, and analyzes data
 - Facilitates **prediction, interpretation, and decision-making**
-

2.1.5 Uses of Statistics

- Drawing representative samples
- Summarizing collected data
- Tabulation and systematic arrangement
- Group comparisons
- Determining behavioral relationships
- Estimating chance vs causation
- Application in:
 - Psychology
 - Education
 - Employment surveys
 - Market Research
 - Industrial and Organizational studies

2.1.6 Limitations of Statistics

- Cannot study **qualitative phenomena** without quantification
- Not applicable to individuals
- **Statistical laws are not exact**
- Does not guarantee **causal relationships**
- Vulnerable to misuse

2.2 Misuse of Statistics

- Use of extremely **small or biased** samples
- **Misleading graphs** or visual misrepresentation
- Illogical or **unexpected comparisons**

Fallacies in Statistics

Fallacies may arise from:

- Poor data collection methods
- Vague or manipulated term definitions
- Improper unit selection

- Faulty classification or grouping
- Inappropriate statistical methods

2.3 Module 2: Mathematics vs Statistics

Aspect	Mathematics	Statistics
Nature	Abstract, symbolic reasoning	Applied, data-based reasoning
Focus	Pure logic, proofs	Real-world data, decision-making
Techniques	Algebra, Calculus, Geometry	Probability, Hypothesis testing, Regression
Output	Theorems, functions, formulas	Inferences, predictions, summaries
Tools	Equations, graphs	Charts, tables, models

2.4 Module 3: Software-Based Statistical Revolution

From Paper to Code

Why shift to software?

- **Faster analysis** of massive data
- **Error-free calculations**
- **Anywhere-anytime** access
- **Cloud-based integration**
- Supports **ML/AI**, automation, and deep visualization

2.4.1 Popular Statistical Software

Software	Type	Use Case
R	Script	Core for academic and professional stats
RKward	GUI	GUI wrapper for R
R Commander	GUI	Menu-based GUI for R
Rattle	GUI	Data mining toolkit in R
Excel	GUI	Basic stats with plugins
Python (pandas)	Script	Modern data science + ML

2.4.2 GUI vs CLI

Feature	GUI (e.g., RKward)	Command Line (e.g., R Console)
Accessibility	User-friendly	Requires learning syntax
Speed	Slower for heavy tasks	High performance
Learning Curve	Minimal	Moderate to High

Feature	GUI (e.g., RKWard)	Command Line (e.g., R Console)
Customization	Limited	Fully scriptable
Teaching Utility	Good for beginners	Good for understanding logic

2.4.3 Recommended GUI Tools for R

- RKWard
- Rattle
- R Commander
- R AnalyticFlow

<https://rkward.kde.org>

2.4.4 Installing RKWard on Ubuntu

bash sudo apt install kbibtex kate libcurl4-openssl-dev libssl-dev libxml2-dev cmake sudo add-apt-repository ppa:rkward-devel/rkward-stable echo "deb https://ppa.launchpad.net/rkward-devel/rkward-stable/ubuntu jammy main" | sudo tee /etc/apt/sources.list.d/rkward.list sudo apt update sudo apt-get install rkward Awesome. Here's Part 2 of the full markdown, Lines 251–600, continuing the structured content from your Week 1 lecture.

2.5 Module 4: Understanding Variables

2.5.1 What is a Variable?

A **variable** is a characteristic or attribute that can assume different values across individuals or items.

In statistics, variables are categorized for analysis and measurement.

2.5.2 R Definition:

In R, variables are containers for data, created by assignment:

```
x <- 10
name <- "Harsh"
flag <- TRUE
```

Classification of Variables

A. Qualitative (Categorical)

Type	Description	Example
------	-------------	---------

Nominal Categories without order `Gender` (Male, Female)
Ordinal Categories with a meaningful order `Education Level` (UG, PG)

B. Quantitative (Numerical)

Type	Description	Example
------	-------------	---------

Discrete	Countable numbers	No. of students
----------	-------------------	-----------------

Continuous	Infinite values in a range	Height, Weight
------------	----------------------------	----------------

Statistical Data Types (Scale of Measurement)

Data Type	Description	Examples
-----------	-------------	----------

Nominal	Categories with no order	Blood group (A, B, AB, O)
---------	--------------------------	---------------------------

Ordinal	Ranked categories	Satisfaction (Low, Med, High)
---------	-------------------	-------------------------------

Interval	Numeric scale with no true zero	Temperature in Celsius
----------	---------------------------------	------------------------

Ratio	Numeric scale with true zero	Income, Weight, Age
-------	------------------------------	---------------------

Data Types in R

R Type	Description	Example	Code
--------	-------------	---------	------

Numeric	Real numbers	<code>x</code>	<code><- 15.3</code>
---------	--------------	----------------	-------------------------

Integer	Whole numbers	<code>y</code>	<code><- as.integer(10)</code>
---------	---------------	----------------	-----------------------------------

Complex	Real + imaginary	<code>z</code>	<code><- 2+3i</code>
---------	------------------	----------------	-------------------------

Character	Text strings	<code>c</code>	<code><- "hello"</code>
-----------	--------------	----------------	----------------------------

Logical	Boolean values	<code>b</code>	<code><- TRUE</code>
---------	----------------	----------------	-------------------------

Factor	Categorical encoding		<code>factor(c("yes", "no", "yes"))</code>
--------	----------------------	--	--

```
# Examples in R
```

```
x <- 15.6
```

```
y <- as.integer(18)
```

```
z <- 7 + 5i
```

```
c <- "I am OK"
```

```
b <- TRUE
```

Module 5: Data Structures in R

Vectors

A vector is a one-dimensional array of elements.

```
vec1 <- c(5, 2, 3, 7, 8, 9, 1, 4, 10, 15)
```

Matrices

Two-dimensional arrays of rows and columns.

```
mat <- matrix(1:9, nrow=3, ncol=3)
```

Arrays

Multidimensional generalization of matrices.

```
arr <- array(1:24, dim=c(3,4,2))
```

Lists

Collection of different types of elements.

```
mylist <- list(name="Alice", age=30, scores=c(89,90))
```

Data Frames

Tabular data (like a spreadsheet), each column can have a different type.

```
df <- data.frame(ID=1:3, Name=c("A", "B", "C"), Score=c(85, 90, 95))
```

Factors

Used for categorical variables.

```
gender <- factor(c("Male", "Female", "Male"))
```

Module 6: Descriptive Statistics

Descriptive statistics summarize and simplify data.

Central Tendency

Measure Formula Meaning

Mean	$\bar{x} = \frac{\sum x_i}{n}$	Average
Median	Middle value in sorted data	Central observation
Mode	Most frequent value	Most common observation

Dispersion Measures

Measure Formula Purpose

Range $\text{Range} = \text{Max} - \text{Min}$ Spread of data

Variance $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$ Spread from mean

Standard Deviation $s = \sqrt{\text{Variance}}$ Average distance from mean

Example in R

```
x <- c(10, 20, 30, 40, 50)
mean(x)
median(x)
var(x)
sd(x)
```

Module 7: Inferential Statistics

Inferential stats allow us to make conclusions about populations using samples.

Key Concepts

Hypothesis Testing: Assesses assumptions about a population.

Confidence Intervals: Estimate population parameters within a range.

Significance Levels (): Commonly 0.05 or 5%

P-Value: Probability of observing the data assuming the null is true.

Hypothesis Types

Type	Description
------	-------------

Null Hypothesis	No difference / no effect
-----------------	---------------------------

Alternative	There is a difference / effect
-------------	--------------------------------

R Examples

```
t.test(x)                             # One-sample t-test
t.test(x, y)                          # Two-sample t-test
```

Module 8: Visualizing Data

Data visualization helps uncover patterns and insights.

Boxplot

Shows 5-number summary

Identifies outliers

```
boxplot(x)
```

Histogram

Frequency distribution of continuous data

```
hist(x)
```

Pie Chart

Shows proportion in categories

```
slices <- c(10, 12, 4, 16, 8)
labels <- c("A", "B", "C", "D", "E")
pie(slices, labels=labels)
```

Scatter Plot

Relationship between two variables

```
plot(x, y)
```

Ogive (Cumulative Frequency)

```
# Create cumulative frequency table manually
```

Module 9: Spreadsheet Basics

Spreadsheets like Excel or Google Sheets are entry points for data work.

Key Features:

Rows → Observations

Columns → Variables

Supports sorting, filtering

Built-in formulas: `=SUM()`, `=AVERAGE()`, etc.

Spreadsheets vs R

Feature `Spreadsheet` (Excel, GSheets) R / RKWard

Cost Usually licensed Free and open source

Flexibility Limited to GUI formulas Full programming capability

Graphics Basic `Advanced` (`ggplot2`)

Reproducibility Low `High` (script-based)

Module `10`: Command Line vs GUI

Command `Line` (R Console)

Windows Command Line

`cd ..`

`mkdir new_folder`

`dir`

R Console Commands

`getwd()`

`setwd("path")`

`install.packages("ggplot2")`

`library(ggplot2)`

GUI (RKWard)

Point-and-click interface

No coding needed

View script history and console

Menu for graphs, models, tables

Learning Resources:

Books

Mohanty, B., & Misra, S. (2016). Statistics for Behavioural and Social Sciences

Pandya et al. (2018). Statistical Analysis in Simple Steps using R

Field, A. P. et al. (2012). Discovering Statistics using R

Harris, J. K. (2019) . Statistics with R: Solving Problems using Real-World Data

Utilizing Statistical Methods for Decision Making

- Use statistical evidence to guide business strategies.
- Make informed policy decisions based on empirical data.
- Report findings clearly for transparency and comprehension.

Summary

The "Basic Statistics Using GUI-R (RK Ward)" course equips learners with the foundational and p

Key Takeaways

- Proficiency in defining and using variables and data types.
- Capability to import and manipulate data in RKWard.
- Understanding of basic statistical practices and their applications.
- Skill in visualizing data for effective communication of results.

Websites

<https://rkward.kde.org>

<https://r4stats.com>

<https://cran.r-project.org>

`<!-- quarto-file-metadata: eyJyZXNvdXJjZURpciI6Ii4ifQ== -->`{=html}

````{=html}

<!-- quarto-file-metadata: eyJyZXNvdXJjZURpciI6Ii4iLCJib29rSXRlbVR5cGUiOiJjaGFwdGVyIiwiaW9va01

## 3 Week 2

### 3.1 Introduction

#### 3.1.1 Purpose of the eBook

This eBook is designed as a complete beginner-to-intermediate guide for understanding the foundational concepts of statistics. It aims to bridge theoretical knowledge and practical application using RKWard (a GUI for R). Readers will be introduced to descriptive and inferential statistics, probability theory, and probability distributions with ample examples and exercises.

#### 3.1.2 Who Should Read This?

- Undergraduate students
- MBA and management students
- Data analysis beginners
- Professionals dealing with data

#### 3.1.3 What You'll Learn

- Data classification and types
  - Descriptive statistics: central tendency and variability
  - Basic probability and events
  - Probability distributions: Bernoulli, Binomial, and Normal
  - Use of RKWard in statistical analysis
- 

## 3.2 1. Fundamentals of Statistics

### 3.2.1 1.1 What is Statistics?

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make informed decisions. It involves both **theoretical** (mathematical) and **applied** approaches to understanding uncertainty and variability in real-world phenomena.

### 3.2.2 1.2 Key Objectives

- Summarizing large datasets effectively
- Estimating population parameters
- Testing hypotheses
- Making predictions and decisions under uncertainty

### 3.2.3 1.3 Types of Statistics

- **Descriptive Statistics:** Deals with the presentation and summarization of data.
  - **Inferential Statistics:** Draws conclusions about populations based on sample data.
- 

## 3.3 2. Types of Data

### 3.3.1 2.1 Classification of Data

| Type         | Example             | Description                |
|--------------|---------------------|----------------------------|
| Qualitative  | Gender, Nationality | Non-numeric labels         |
| Quantitative | Height, Age         | Numeric values             |
| Discrete     | No. of Children     | Countable numbers          |
| Continuous   | Temperature, Weight | Infinite values in a range |

#### 3.3.1.1 Qualitative (Categorical) Data

- **Nominal:** No inherent order (e.g., religion, marital status).
- **Ordinal:** Natural order (e.g., customer satisfaction: Poor, Average, Good).

#### 3.3.1.2 Quantitative (Numerical) Data

- **Discrete:** Integers; e.g., number of books.
  - **Continuous:** Measurable; e.g., weight in kilograms.
-



## 3.4 3. Descriptive Statistics

### 3.4.1 3.1 Measures of Central Tendency

#### 3.4.1.1 What is Central Tendency?

Central tendency refers to the center or middle of a dataset. It's the value that best represents the entire distribution.

#### 3.4.1.2 Characteristics of a Good Measure

- Rigidly defined
  - Easy to understand
  - Takes all data into account
  - Amenable to algebraic treatment
  - Stable under sampling
  - Minimally affected by outliers (except mean)
- 

### 3.4.2 3.2 The Mean

#### 3.4.2.1 Definition

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

#### 3.4.2.2 Formula

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

#### 3.4.2.3 Properties of Mean

- Uses all data values
- Affected by extreme values
- The sum of deviations from the mean is zero

#### 3.4.2.4 Example

Data: 10, 15, 20, 25, 30

Mean =  $(10 + 15 + 20 + 25 + 30)/5 = 20$

---

### **3.4.3 3.3 The Median**

#### **3.4.3.1 Definition**

The median is the value separating the higher half from the lower half of a data sample.

#### **3.4.3.2 Calculation**

- Odd number of items: Middle value
- Even number of items: Average of the two middle values

#### **3.4.3.3 Properties**

- Not influenced by extreme values
- Best for skewed data

#### **3.4.3.4 Example**

Data: 4, 6, 9, 12, 15, 21, 33

Median = 12 (middle value)

---

### **3.4.4 3.4 The Mode**

#### **3.4.4.1 Definition**

The mode is the value that appears most frequently in a dataset.

#### **3.4.4.2 Characteristics**

- Can be used for categorical data
- Dataset can be unimodal, bimodal, or multimodal
- May not exist if all values are unique

#### **3.4.4.3 Example**

Data: 4, 4, 6, 8, 9, 10, 4

Mode = 4

---

### 3.4.5 3.5 Comparison Table

| Measure | Use Case                | Affected by Outliers | Mathematical Use |
|---------|-------------------------|----------------------|------------------|
| Mean    | Symmetric distributions | Yes                  | High             |
| Median  | Skewed distributions    | No                   | Moderate         |
| Mode    | Categorical variables   | No                   | Low              |

---

## 3.5 4. Measures of Variability

### 3.5.1 4.1 Why Measure Variability?

While central tendency summarizes data, variability tells us how spread out the data is. It's essential in determining consistency and reliability.

---

### 3.5.2 4.2 Range

#### 3.5.2.1 Definition

The difference between the maximum and minimum values.

$$\text{Range} = x_{\max} - x_{\min}$$

#### 3.5.2.2 Example

Data: 12, 14, 17, 19, 23

$$\text{Range} = 23 - 12 = 11$$

#### 3.5.2.3 Limitations

- Ignores distribution shape
- Extremely sensitive to outliers

### 3.5.3 4.3 Quartiles and Interquartile Range

#### 3.5.3.1 Quartiles

- Q1 (25th percentile): Lower quartile
- Q2 (50th percentile): Median
- Q3 (75th percentile): Upper quartile

#### 3.5.3.2 Formula for Position

$$Q_k = \frac{k(n+1)}{4}$$

#### 3.5.3.3 IQR Formula

$$IQR = Q3 - Q1$$

#### 3.5.3.4 Example

Data: 12, 30, 45, 57, 70

Q1 = 30, Q3 = 57 → IQR = 27

---

### 3.5.4 4.4 Variance

#### 3.5.4.1 Concept

Variance is the average of the squared differences from the Mean.

#### 3.5.4.2 Formulas

Population Variance:

$$\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$$

Sample Variance:

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

---

### 3.5.5 4.5 Standard Deviation

#### 3.5.5.1 Concept

Standard deviation is the square root of variance. It provides a measure of spread in the same units as the data.

$$s = \sqrt{s^2}$$

#### 3.5.5.2 Properties

- Same unit as original data
  - Measures how far values deviate from the mean
  - Widely used in most statistical computations
- 

### 3.5.6 4.6 Coefficient of Variation (CV)

#### 3.5.6.1 Definition

The ratio of the standard deviation to the mean, expressed as a percentage. Used to compare variability between datasets with different units.

$$CV = \left( \frac{s}{\bar{x}} \right) \times 100\%$$

#### 3.5.6.2 Example

Dataset A: Mean = 100, SD = 10  $\rightarrow$  CV = 10%

Dataset B: Mean = 50, SD = 5  $\rightarrow$  CV = 10%

---

### 3.5.7 4.7 Moment-Based Measures

- First Moment (about mean): 0 (since  $\sum (x - \bar{x}) = 0$ )
  - Second Moment: Variance
  - Third Moment: Skewness
  - Fourth Moment: Kurtosis
-

## 3.6 5. Probability Fundamentals

### 3.6.1 5.1 Introduction to Probability

Probability is the mathematical framework for quantifying uncertainty. It helps us estimate how likely an event is to occur.

### 3.6.2 5.2 Key Definitions

- **Experiment:** A process that leads to an outcome.
  - **Outcome:** The result of an experiment.
  - **Sample Space ( $\Omega$ ):** All possible outcomes.
  - **Event:** A subset of the sample space.
- 

### 3.6.3 5.3 Types of Events

| Event Type         | Description                                 |
|--------------------|---------------------------------------------|
| Independent        | Occurrence of one does not affect the other |
| Dependent          | One affects the outcome of another          |
| Mutually Exclusive | Cannot occur together                       |
| Exhaustive         | Includes all possible outcomes              |

---

### 3.6.4 5.4 Classical Probability

Used when all outcomes are equally likely.

**Formula:**

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total outcomes in } \Omega}$$

**Example:** Rolling a fair die

$$P(\text{rolling a 3}) = 1/6$$

---

### 3.6.5 5.5 Probability Rules

#### 3.6.5.1 Rule 1: Non-Negativity

$$0 \leq P(A) \leq 1$$

### 3.6.5.2 Rule 2: Total Probability

$$P(\Omega) = 1$$

### 3.6.5.3 Rule 3: Complement Rule

$$P(A^c) = 1 - P(A)$$

### 3.6.5.4 Rule 4: Addition Rule

If A and B are mutually exclusive:

$$P(A \cup B) = P(A) + P(B)$$

Otherwise:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

### 3.6.5.5 Rule 5: Multiplication Rule

- For independent events:

$$P(A \cap B) = P(A) \cdot P(B)$$

---

## 3.6.6 5.6 Conditional Probability

Formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

---

## 3.7 6. Discrete Probability Distributions

### 3.7.1 6.1 Bernoulli Distribution

- One trial, two outcomes (success/failure).
- Success = 1, Failure = 0

$$P(X = x) = p^x(1 - p)^{1-x}, \quad x \in \{0, 1\}$$

- Mean =  $p$
- Variance =  $p(1 - p)$

#### 3.7.1.1 Example:

Flip a fair coin  $\rightarrow p = 0.5$

Mean = 0.5, Variance = 0.25

---

### 3.7.2 6.2 Binomial Distribution

- Series of  $n$  independent Bernoulli trials
- Number of successes  $x$  out of  $n$  trials

**Formula:**

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- Mean:  $\mu = np$
- Variance:  $\sigma^2 = np(1 - p)$

#### 3.7.2.1 Example:

Flip a coin 5 times ( $p = 0.5$ )

$$P(X = 3) = \binom{5}{3} (0.5)^3 (0.5)^2 = 10 \cdot 0.125 \cdot 0.25 = 0.3125$$

---



## 3.8 7. Continuous Distributions

### 3.8.1 7.1 Normal Distribution

The most important continuous distribution in statistics.

**Properties:**

- Bell-shaped and symmetric
- Defined by mean (  $\mu$  ) and variance (  $\sigma^2$  )
- Total area under the curve = 1

**Probability Density Function (PDF):**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

#### 3.8.1.1 Empirical Rule:

- 68% of values lie within  $\pm 1$
  - 95% within  $\pm 2$
  - 99.7% within  $\pm 3$
- 

### 3.8.2 7.2 Standard Normal Distribution

A normal distribution with:

- Mean = 0
- Standard deviation = 1

**Z-score Formula:**

$$Z = \frac{X - \mu}{\sigma}$$

#### 3.8.2.1 Example:

If  $\mu = 100$ ,  $\sigma = 15$ , and  $X = 130$

Then  $Z = \frac{130-100}{15} = 2$

---

## 3.9 8. Visualizing Data

### 3.9.1 8.1 Frequency Distribution

| Class Interval | Frequency |
|----------------|-----------|
| 0–10           | 3         |
| 11–20          | 7         |
| 21–30          | 9         |
| 31–40          | 6         |

---

### 3.9.2 8.2 Histogram

A bar chart representing the frequency distribution of numerical data.

**Use Case:** Visualize shape (e.g., normal, skewed)

---

### 3.9.3 8.3 Boxplot (Box-and-Whisker Plot)

Shows:

- Minimum
- Q1
- Median
- Q3
- Maximum
- Outliers (as dots)

Helps identify skewness and outliers quickly.

---

### 3.9.4 8.4 Scatter Plot

Used to study the relationship between two quantitative variables.

---

## 3.10 9. Practical Applications

### 3.10.1 9.1 Business Use Cases

- Retail: Analyze sales patterns
  - Healthcare: Patient outcome probabilities
  - Finance: Stock volatility (using SD, CV)
- 

### 3.10.2 9.2 Education and Research

- Student test scores: Use mean, SD, and percentile ranking
  - Experiment analysis: Use Z-scores and Normal Distribution
- 

## 3.11 10. Using RKWard

### 3.11.1 10.1 What is RKWard?

A graphical frontend for the R programming language designed for statistical analysis and data visualization.

---

### 3.11.2 10.2 Installation Guide

1. Download R from [CRAN](https://cran.r-project.org/)
  2. Install RKWard from [rkward.kde.org](https://rkward.kde.org/)
  3. Start RKWard and begin with menu-driven tasks
- 

### 3.11.3 10.3 Sample RKWard Activities

#### 3.11.3.1 Calculate Mean and SD

- Load dataset
- Click *Statistics* → *Descriptive Statistics*
- Choose variables and click *OK*

### 3.11.3.2 Visualize Histogram

- Click *Graphics* → *Histogram*
- Select variable and customize bins

---

### 3.11.4 10.4 Using R Code in RKWard

```
data <- c(12, 15, 17, 18, 21)
mean(data)
sd(data)
hist(data)
```

*## Summary*

This eBook provided a deep dive into basic statistics including:

Data types and classification  
Central tendency and variability  
Probability theory and rules  
Discrete and continuous distributions  
Visual interpretation and real-world applications  
GUI-based statistical analysis using RKWard

```
`<!-- quarto-file-metadata: eyJyZXNvdXJjZURpciI6Ii4ifQ== -->`{=html}
```

```
````{=html}
```

```
<!-- quarto-file-metadata: eyJyZXNvdXJjZURpciI6Ii4iLCJib29rSXRlbVR5cGUiOiJjaGFwdGVyIiwiaW9va01
```

4 Week 3

5 Introduction

5.1 Importance of Statistics

Statistics is a powerful tool used across disciplines — from economics and psychology to biology, data science, and machine learning. It enables:

- Interpretation of data
- Generalization from samples to populations
- Hypothesis testing and decision-making
- Prediction and modeling

Understanding statistics is essential for anyone involved in **empirical research**, **policy making**, **data-driven decision-making**, or **scientific inquiry**.

5.2 Overview of Topics

This book covers:

- Population vs Sample
 - Hypotheses and Errors
 - Descriptive vs Inferential Statistics
 - Data Types (R + Theoretical)
 - Sampling Techniques
 - Normal Distribution
 - Linear and Logistic Regression
 - GUI-based R interfaces: RKWard, Rcmdr, Rattle
 - Fallacies and misuse in statistics
 - Graphical Methods
 - R programming constructs for statistics
-

6 Understanding Populations and Samples

6.1 Population

The complete set of all units of interest. Examples:

- All students in India
- All electric cars in the U.S.

6.2 Sample

A **subset of the population**, selected for analysis. Goal: represent the population accurately.

6.3 Why Use Samples?

- More practical and cost-efficient
- Enables faster analysis
- Allows estimation and inference

6.4 Relation Between Population & Sample

Population → Sample → Statistic → Inference → Population Parameter

7 Hypotheses and Errors

7.1 Hypothesis Defined

A hypothesis is a testable assumption about a population.

7.1.1 Null Hypothesis (H_0)

- No difference or effect
- Example: H_0 : “ = 100”

7.1.2 Alternative Hypothesis (H_A)

- A difference or effect exists
- Example: H_A : “ ≠ 100”

7.2 Types of Errors

| Error Type | Description |
|---------------|---|
| Type I Error | Rejecting H_0 when it's true (false positive) |
| Type II Error | Failing to reject H_0 when it's false (false neg) |

7.3 Significance Level ()

The probability of making a Type I error — commonly set to **0.05 (5%)**

8 Inferential Statistics

8.1 Purpose

- Estimate unknown population parameters
- Test hypotheses
- Predict outcomes

8.2 Common Techniques

- t-test
 - z-test
 - ANOVA
 - Chi-square
 - Regression
-

8.3 Sampling Techniques

8.3.1 1. Simple Random Sampling

Every unit has equal probability.

8.3.2 2. Systematic Sampling

Pick every k th element.

8.3.3 3. Stratified Sampling

Subdivide population into strata (e.g. age groups), then sample from each.

8.3.4 4. Cluster Sampling

Randomly choose entire groups (e.g. schools, cities).

8.4 Central Limit Theorem (CLT)

If $n > 30$, the distribution of sample means approximates a **normal distribution** even if the original population is not normal.

Formula:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

9 Descriptive Statistics

9.1 Measures of Central Tendency

9.1.1 Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

9.1.2 Median

Middle value in an ordered dataset.

9.1.3 Mode

Most frequent value.

9.2 Measures of Dispersion

9.2.1 Range

$$Range = Max - Min$$

9.2.2 Variance

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

9.2.3 Standard Deviation

$$s = \sqrt{s^2}$$

9.3 Measures of Shape

- **Skewness:** Degree of asymmetry
 - **Kurtosis:** Peakedness of distribution
-

10 Graphical Methods

10.1 Histogram

`hist(data$height, col="blue", main="Height Distribution")` Boxplot

`boxplot(data$score, data$group)` Scatter Plot

`plot(data$x, data$y, col="red")` Ogive (Cumulative Frequency Plot)

Built using cumulative frequency of class intervals.

10.2 R Data Types and Structures

Basic Data Types

`x <- 12.5` # numeric `y <- as.integer(5)` # integer `z <- 4 + 3i` # complex `name <- "Ravi"` # character `flag <- TRUE` # logical Vectors

`v <- c(1, 2, 3)` Matrices

`m <- matrix(1:9, nrow=3, byrow=TRUE)` Data Frame

`df <- data.frame(Name=c("A", "B"), Score=c(89, 94))` Lists

`lst <- list(id=101, name="John", marks=c(78, 82))` Factors

`gender <- factor(c("Male", "Female", "Male"))` Statistical Fallacies

What are Fallacies?

Fallacies occur when conclusions are drawn based on flawed statistical reasoning.

Common Fallacies

Improper Sampling Misleading Graphs Ambiguous Term Definitions Ignoring Confounding Variables Assuming Correlation Implies Causation Misuse of Statistics

Examples of Misuse

Using biased samples Cherry-picking data Using 3D pie charts to exaggerate results Misrepresenting scale in graphs

10.3 Comparing R vs Excel vs GUI-R (RKWard)

| Feature | R (Script) | Excel | RKWard GUI |
|-------------------|------------|------------|-------------------|
| Usability | Medium | Easy | Easy |
| Flexibility | High | Low-Medium | Medium |
| Statistical Power | Very High | Low | High |
| Graphics | ggplot2 | Basic | ggplot2 supported |
| Reproducibility | High | Low | High |

10.4 Installing RKWard (Ubuntu)

```
sudo apt install kbbibtex kate libcurl4-openssl-dev libssl-dev libxml2-dev cmake sudo add-apt-repository ppa:rkward
```

10.5 Teaching Tools in RKWard

```
install.packages(c("R2HTML", "car", "e1071", "Hmisc", "plyr", "ggplot2", "prob", "ez", "multcomp", "remotes"), de
```

10.6 GUI-Based Statistical Tools

RKWard – KDE interface for R
 Rcmdr – Classic R Commander GUI
 Rattle – Data mining GUI
 in R
 R AnalyticFlow – Flow-based programming for statistics

11 Linear Regression in R

11.1 What is Linear Regression?

Linear regression models the relationship between a **dependent variable (Y)** and one or more **independent variables (X)**.

11.1.1 Simple Linear Regression Equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y is the dependent variable
- X is the independent variable
- β_0 is the intercept
- β_1 is the slope
- ϵ is the error term

11.2 Code Example

```
r # Load data data(mtcars)
```

12 Fit model

```
model <- lm(mpg ~ wt, data=mtcars)
```


13 Summary

`summary(model)`

13.1 Adjusted R-squared

Penalizes the number of predictors to avoid overfitting.

AIC & BIC

AIC: Akaike Information Criterion BIC: Bayesian Information Criterion Lower values of AIC/BIC
→ better model fit (with penalty for complexity).

13.2 Normal Distribution

Key Properties

Symmetrical, bell-shaped curve Mean = Median = Mode Total area under curve = 1 Empirical Rule: 68% within ± 1 SD 95% within ± 2 SD 99.7% within ± 3 SD

Example: Given: Mean = 70, SD = 5, X = 75

`z <- (75 - 70) / 5` # Result: 1.0 Z-Table Usage

Find the area under the curve to the left of the z-score Useful for probability and percentile ranking

13.3 Data Import Techniques

CSV Import in R

`df <- read.csv("data.csv", header=TRUE)` head(df) Excel Import (using readxl)

`install.packages("readxl")` library(readxl)

`df <- read_excel("data.xlsx")`

13.4 Working with the RKWard Interface

Sections: Console – Run R code Script Editor – Write reusable code Workspace – View loaded variables Teaching Tab – Education-focused modules

13.5 Spreadsheet Concepts

Structure

Component | Description Rows | Individual observations Columns | Variables Cells | Data points
Header Row | Variable names

13.6 Advantages

Easy data entry Visual inspection Good for small datasets

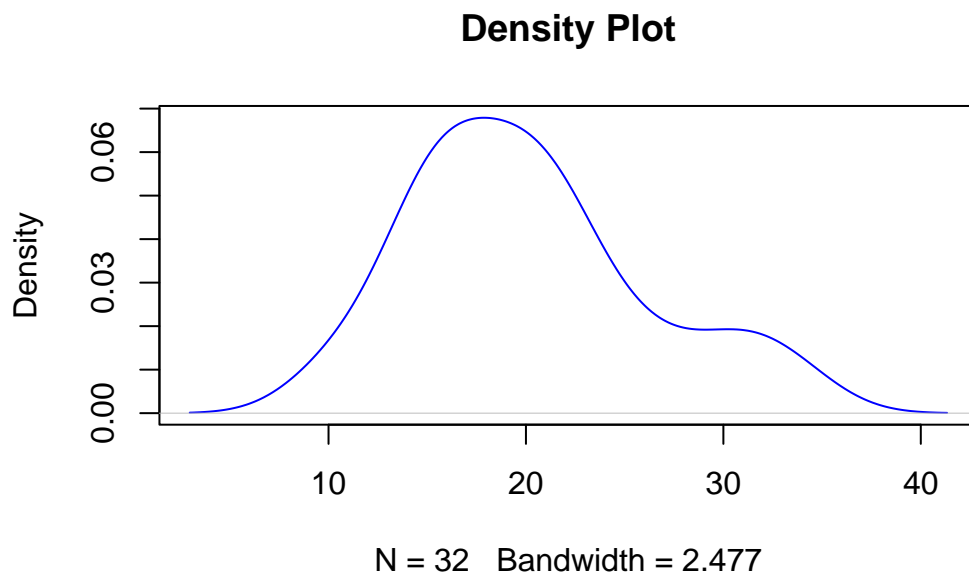
13.7 Limitations

Limited statistical functionality Hard to reproduce Error-prone for large datasets

13.8 Advanced Plots and Techniques

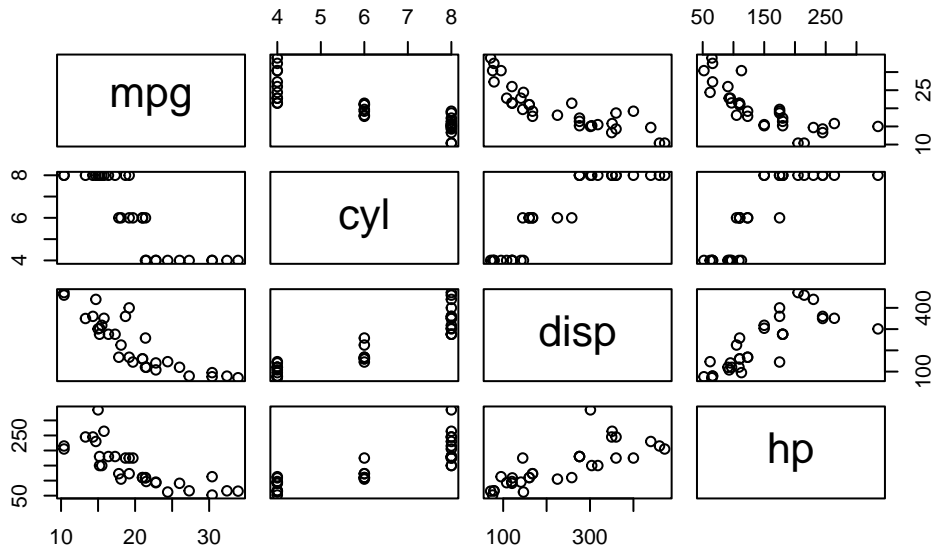
Density Plot

```
plot(density(mtcars$mpg), main="Density Plot", col="blue")
```



Pair Plot

```
pairs(mtcars[, 1:4])
```



Correlation Matrix

```
cor(mtcars)
```

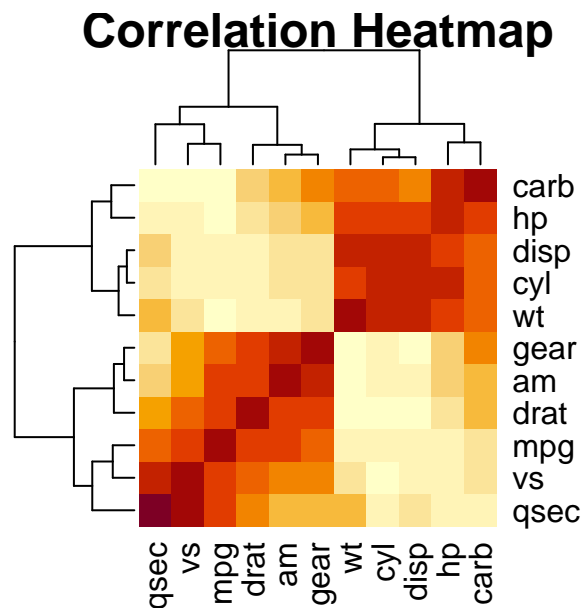
| | mpg | cyl | disp | hp | drat | wt |
|------|------------|------------|------------|------------|-------------|------------|
| mpg | 1.0000000 | -0.8521620 | -0.8475514 | -0.7761684 | 0.68117191 | -0.8676594 |
| cyl | -0.8521620 | 1.0000000 | 0.9020329 | 0.8324475 | -0.69993811 | 0.7824958 |
| disp | -0.8475514 | 0.9020329 | 1.0000000 | 0.7909486 | -0.71021393 | 0.8879799 |
| hp | -0.7761684 | 0.8324475 | 0.7909486 | 1.0000000 | -0.44875912 | 0.6587479 |
| drat | 0.6811719 | -0.6999381 | -0.7102139 | -0.4487591 | 1.00000000 | -0.7124406 |
| wt | -0.8676594 | 0.7824958 | 0.8879799 | 0.6587479 | -0.71244065 | 1.0000000 |
| qsec | 0.4186840 | -0.5912421 | -0.4336979 | -0.7082234 | 0.09120476 | -0.1747159 |
| vs | 0.6640389 | -0.8108118 | -0.7104159 | -0.7230967 | 0.44027846 | -0.5549157 |
| am | 0.5998324 | -0.5226070 | -0.5912270 | -0.2432043 | 0.71271113 | -0.6924953 |
| gear | 0.4802848 | -0.4926866 | -0.5555692 | -0.1257043 | 0.69961013 | -0.5832870 |
| carb | -0.5509251 | 0.5269883 | 0.3949769 | 0.7498125 | -0.09078980 | 0.4276059 |

| | qsec | vs | am | gear | carb |
|------|-------------|------------|-------------|------------|-------------|
| mpg | 0.41868403 | 0.6640389 | 0.59983243 | 0.4802848 | -0.55092507 |
| cyl | -0.59124207 | -0.8108118 | -0.52260705 | -0.4926866 | 0.52698829 |
| disp | -0.43369788 | -0.7104159 | -0.59122704 | -0.5555692 | 0.39497686 |
| hp | -0.70822339 | -0.7230967 | -0.24320426 | -0.1257043 | 0.74981247 |
| drat | 0.09120476 | 0.4402785 | 0.71271113 | 0.6996101 | -0.09078980 |
| wt | -0.17471588 | -0.5549157 | -0.69249526 | -0.5832870 | 0.42760594 |
| qsec | 1.00000000 | 0.7445354 | -0.22986086 | -0.2126822 | -0.65624923 |
| vs | 0.74453544 | 1.0000000 | 0.16834512 | 0.2060233 | -0.56960714 |

| | | | | | |
|------|-------------|------------|------------|------------|------------|
| am | -0.22986086 | 0.1683451 | 1.00000000 | 0.7940588 | 0.05753435 |
| gear | -0.21268223 | 0.2060233 | 0.79405876 | 1.00000000 | 0.27407284 |
| carb | -0.65624923 | -0.5696071 | 0.05753435 | 0.2740728 | 1.00000000 |

Heatmap

```
heatmap(cor(mtcars), main="Correlation Heatmap")
```



13.9 Common R Packages for Statistics

Package | Purpose ggplot2 | Data visualization dplyr | Data manipulation tidyr | Data tidying Hmisc
| Misc stats functions car | Regression diagnostics e1071 | Skewness/kurtosis, ML tools psych | Psy-
chological statistics shiny | Interactive apps caret | Classification and regression

13.10 Introduction to Command Line

13.11 Windows Terminal

```
cd..mkdirmy_projectdir
```

Linux Terminal

```
cd mkdirstats_projectls -l
```

13.12 Git + R Project Example

*git init
git clone https://github.com/username/project.git*

13.13 R Script Template for Analysis

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

14 Load data

```
df <- read.csv("dataset.csv")
```

15 Descriptive stats

```
summary(df) sd(df$Score)
```

16 Histogram

```
ggplot(df, aes(x=Score)) + geom_histogram(bins=10, fill="skyblue")
```


17 Linear regression

```
model <- lm(Score ~ StudyHours, data=df) summary(model)
```

18 Scatter plot with regression line

```
ggplot(df, aes(x=StudyHours, y=Score)) + geom_point() + geom_smooth(method="lm") $$ ##  
Fallacies and Bias: Real-World Cautions
```

Examples of Statistical Abuse

Cherry-picking data Data dredging (p-hacking) Using relative risk without absolute context Non-random sampling Ethics in Data Analysis

Be transparent Document sources Disclose methodology Avoid overstating conclusions

18.1 Future Applications of Statistics

Real-World Domains

Healthcare: Drug effectiveness, diagnostics Economics: Forecasting, policy evaluation Sociology: Survey analysis Sports: Performance analytics AI/ML: Predictive modeling, optimization Next Steps

Learn tidyverse ecosystem Explore machine learning in R Build Shiny dashboards Get familiar with reproducible research using Quarto

18.2 Practice Challenges

1. Load and summarize data

Load mtcars or your own dataset Use summary(), mean(), sd() 2. Create 3 different plots

Histogram Boxplot by group Scatter plot with trend line 3. Build a regression model

Identify predictor and outcome Use lm() and summary() 4. Explore a GUI like RKWard or Rcmdr

18.3 Key Takeaways

Statistics supports informed decision-making. R and its GUI frontends offer flexibility + power. Understand theory → then automate with code. Avoid fallacies by following robust methods. Visuals are crucial: plot early, plot often.

19 Week 4

20 Introduction

This eBook is a comprehensive companion to the course *Basic Statistics using GUI-R (RKWard)*. It includes foundational theory, practical examples, and step-by-step explanations, with integrated GUI-R usage.

21 Course Overview

21.1 Course Name

Basic Statistics using GUI-R (RKWard)

21.2 Instructor Profile

Dr. Harsh Pradhan is Assistant Professor at the Institute of Management Studies, Banaras Hindu University.

[Faculty Profile](#)

21.3 Learning Objectives

- Understand core concepts in statistics
- Apply t-tests and ANOVA using real data
- Compute confidence intervals and test statistics
- Use GUI-R (RKWard) for statistical analysis

22 Chapter 1: Fundamental Concepts

22.1 Descriptive Statistics

22.1.1 Central Tendency

- Mean
- Median
- Mode

22.1.2 Dispersion

- Range
- Variance
- Standard Deviation

22.1.2.1 Example:

```
data <- c(4, 8, 6, 5, 3)
mean(data)
```

```
[1] 5.2
```

```
median(data)
```

```
[1] 5
```

```
sd(data)
```

```
[1] 1.923538
```

22.2 Standard Error

$$SE = \frac{s}{\sqrt{n}}$$

Small SE = sample mean is a good estimate of the population mean.

22.3 Central Limit Theorem

For $n > 30$, sampling distribution of the mean approximates normal:

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

22.4 Confidence Intervals

$$CI = \bar{x} \pm Z \cdot \frac{s}{\sqrt{n}}$$

Interpret 95% CI as: 95 of 100 such intervals would contain the true mean.

23 Chapter 2: Estimation

23.1 Types of Estimates

| Type | Description | Example |
|-------------------|--------------------|----------------|
| Point Estimate | Single value | Sample mean |
| Interval Estimate | Range + confidence | Confidence Int |

23.2 Parameter vs Statistic

| Term | Description |
|-----------|--------------------------------------|
| Parameter | Value from population (e.g., μ) |
| Statistic | Value from sample (e.g., \bar{x}) |

24 Chapter 3: Hypothesis Testing

- **Null Hypothesis (H_0):** No effect
- **Alternative Hypothesis (H_1):** Some effect
- **Type I Error:** Reject H_0 when true
- **Type II Error:** Fail to reject H_0 when false

25 Chapter 4: Student's T-Test

25.1 Types

| Test Type | Description |
|-------------|-------------------------------|
| One-Sample | Compare sample to fixed value |
| Independent | Compare two unrelated groups |
| Paired | Compare two related groups |

25.2 One-Sample T-Test Example

```
data <- c(22, 24, 27, 26, 28, 23, 25, 29, 21, 26, 24, 27)
t.test(data, mu = 25)
```

One Sample t-test

```
data: data
t = 0.2363, df = 11, p-value = 0.8175
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 23.61427 26.71906
sample estimates:
mean of x
 25.16667
```

25.3 Test Statistic

$$t = \frac{\bar{x} - \mu}{SE}$$

25.4 Degrees of Freedom

$$df = n - 1$$

25.5 Decision Rule

Compare calculated t to table value. If $|t| > t_{critical}$, reject H_0 .

25.6 T-Test in GUI-R

1. Import data
2. Choose T-Test
3. Define groups
4. Run & interpret output

26 Chapter 5: ANOVA

26.1 Purpose

Used when comparing means across 3+ groups.

26.1.1 One-Way ANOVA Formula

$$F = \frac{MS_{between}}{MS_{within}}$$

Where:

- $MS_{between} = \frac{SS_{between}}{df_{between}}$
- $MS_{within} = \frac{SS_{within}}{df_{within}}$

26.1.2 Assumptions

- Normality
- Homogeneity of variance
- Independence

26.1.3 Example Table

| Group | Mean | Var | n |
|-------|------|-----|----|
| A | 5.5 | 1.5 | 30 |
| B | 7.1 | 2.0 | 30 |
| C | 6.8 | 1.8 | 30 |

26.2 Post-Hoc Tests

Run if ANOVA is significant to locate pairwise differences.

26.3 ANOVA in GUI-R

1. Load data
2. Choose “One-Way ANOVA”
3. Define groups
4. Interpret output

27 Chapter 6: GUI-R Workflow

1. **Import Data** (CSV, Excel)
2. **Choose Test** (T-Test, ANOVA, etc.)
3. **Run** the analysis
4. **Interpret** the output
5. **Export** the results or visualizations

28 Chapter 7: Advanced Concepts

28.1 Variance Partitioning

$$\text{Total Variance} = \text{Explained Variance} + \text{Unexplained Variance}$$

| Explained Terms | Unexplained Terms |
|-----------------|-------------------|
| Systematic | Random |
| Predictive | Error |
| Deterministic | Noise |

28.2 Degrees of Freedom

For equation $x + y + z = 3$, if 2 values are known, third is fixed.
Hence, $df = n - k$ where n = total variables, k = constraints.

28.3 Chi-Square and F Distribution

- **Chi-Square:** Categorical variable comparison
- **F-Distribution:** Used in ANOVA, variance testing

28.4 Univariate, Bivariate, Multivariate

| Type | Variables | Example |
|--------------|-----------|------------------------------|
| Univariate | 1 | Height |
| Bivariate | 2 | Height vs Weight |
| Multivariate | >2 | Study w/ Age, Gender, Income |

28.5 Parametric Test Assumptions

- Interval/Ratio DV
- Random Sampling
- Normality
- Equal Variances

If assumptions violated → use non-parametric test.

28.6 Effect Size

$$\text{Effect Size} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Used for comparison across studies.

28.7 Power of a Test

$$\text{Power} = 1 - \beta$$

Higher power → lower chance of Type II error

Power increases with sample size, effect size

29 Conclusion

Statistics is the language of data. GUI-R makes statistical tools accessible for everyone. This book empowers you to analyze data effectively using t-tests, ANOVA, and confidence intervals in a GUI environment.

30 References

- Pradhan, H. (2023). *Basic Statistics using GUI-R (RKWard)*
- Field, A. (2013). *Discovering Statistics Using R*.
- <https://methods.sagepub.com>

31 Chapter 8: Advanced T-Test Applications

31.1 Paired Sample T-Test

Used when the same group is measured twice (e.g., before and after).

31.1.1 Example:

```
before <- c(80, 82, 79, 84, 88)
after <- c(78, 81, 76, 83, 86)
t.test(before, after, paired = TRUE)
```

Paired t-test

```
data: before and after
t = 4.8107, df = 4, p-value = 0.008581
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.7611494 2.8388506
sample estimates:
mean difference
      1.8
```

31.2 Independent Samples T-Test

Compare means of two unrelated groups.

```
group1 <- c(85, 90, 88, 92, 87)
group2 <- c(80, 83, 85, 84, 82)
t.test(group1, group2)
```

Welch Two Sample t-test

```
data: group1 and group2
t = 3.7755, df = 7.226, p-value = 0.006537
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.114814 9.085186
sample estimates:
mean of x mean of y
 88.4      82.8
```

31.3 One-Sample T-Test with GUI-R

- Import dataset
- Use 'Descriptive Statistics' to check mean
- Navigate to 'T-Test' → 'One Sample'
- Input hypothesized mean and run

32 Chapter 9: More on Confidence Intervals

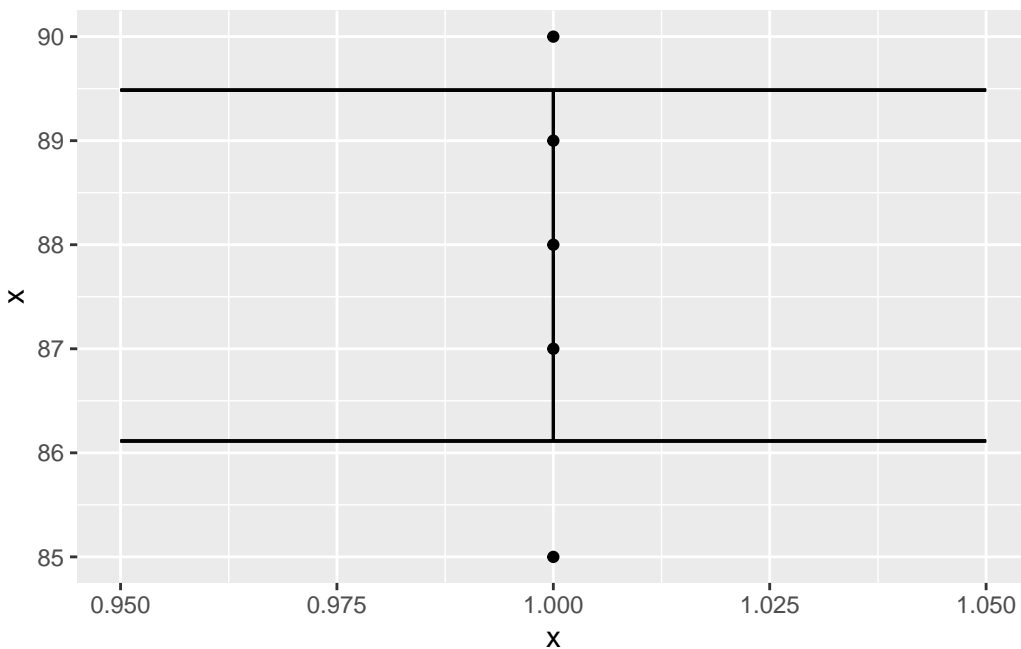
32.1 Visualizing Confidence Intervals in R

```
x <- c(88, 90, 85, 87, 89)
mean_x <- mean(x)
se <- sd(x) / sqrt(length(x))
ci_lower <- mean_x - 1.96 * se
ci_upper <- mean_x + 1.96 * se
c(ci_lower, ci_upper)
```

```
[1] 86.11394 89.48606
```

Plot using ggplot2:

```
library(ggplot2)
df <- data.frame(x = x)
ggplot(df, aes(y = x, x = 1)) +
  geom_point() +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), width = 0.1)
```



33 Chapter 10: Robust ANOVA Models

33.1 Two-Way ANOVA

Examines the effect of two categorical independent variables on a continuous dependent variable.

```
# Sample dataset for demonstration
dataset <- data.frame(
  score = c(85, 90, 88, 92, 87, 80, 83, 85, 84, 82),
  gender = rep(c("Male", "Female"), each = 5),
  teaching_method = rep(c("A", "B"), times = 5)
)
aov_result <- aov(score ~ gender * teaching_method, data = dataset)
summary(aov_result)
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|------------------------|----|--------|---------|---------|------------|
| gender | 1 | 78.40 | 78.40 | 23.718 | 0.00279 ** |
| teaching_method | 1 | 6.02 | 6.02 | 1.820 | 0.22598 |
| gender:teaching_method | 1 | 18.15 | 18.15 | 5.491 | 0.05759 . |
| Residuals | 6 | 19.83 | 3.31 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

33.2 Repeated Measures ANOVA

Use when the same subjects are used for each treatment.

```
# Sample repeated measures data in long format
data_long <- data.frame(
  id = rep(1:5, each = 3),
  condition = rep(c("A", "B", "C"), times = 5),
  score = c(85, 88, 90, 80, 82, 85, 78, 80, 83, 90, 92, 95, 88, 90, 91)
)
library(ez)
ezANOVA(data = data_long, dv = .(score), wid = .(id), within = .(condition))
```

Warning: Converting "id" to factor for ANOVA.

Warning: Converting "condition" to factor for ANOVA.

\$ANOVA

| | Effect | DFn | DFd | F | p | p<.05 | ges |
|---|-----------|-----|-----|----------|--------------|-------|-----------|
| 2 | condition | 2 | 8 | 88.22222 | 3.539139e-06 | * | 0.1479687 |

\$`Mauchly's Test for Sphericity`

| | Effect | W | p | p<.05 |
|---|-----------|-----------|-----------|-------|
| 2 | condition | 0.5555556 | 0.4140867 | |

\$`Sphericity Corrections`

| | Effect | GGe | p[GG] | p[GG]<.05 | HFe | p[HF] | p[HF]<.05 |
|---|-----------|-----------|--------------|-----------|-----------|--------------|-----------|
| 2 | condition | 0.6923077 | 9.135419e-05 | * | 0.9411765 | 6.568851e-06 | * |

34 Chapter 11: Effect Size Measures

34.1 Cohen's d

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

Where s_p is the pooled standard deviation.

34.1.1 R Example

```
library(effsize)
cohen.d(group1, group2)
```

Cohen's d

```
d estimate: 2.387848 (large)
95 percent confidence interval:
      lower      upper
0.4791634 4.2965327
```

34.2 Eta-Squared (η^2)

Used for ANOVA:

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

35 Chapter 12: Statistical Assumptions Checking

35.1 Normality

Use Shapiro-Wilk test:

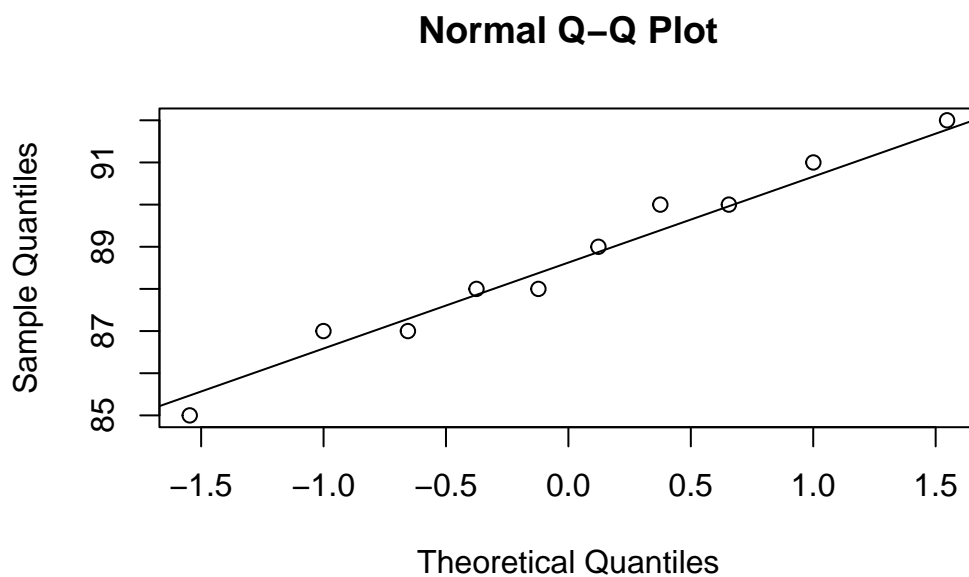
```
# Sample data frame for normality test
data <- data.frame(variable = c(88, 90, 85, 87, 89, 91, 92, 88, 90, 87))
shapiro.test(data$variable)
```

Shapiro-Wilk normality test

```
data: data$variable
W = 0.97743, p-value = 0.95
```

Visualize:

```
qqnorm(data$variable)
qqline(data$variable)
```



35.2 Homogeneity of Variance

Use Levene's Test:

```
# Sample data frame for Levene's Test
data <- data.frame(
  variable = c(88, 90, 85, 87, 89, 91, 92, 88, 90, 87),
  group = rep(c("A", "B"), each = 5)
)
library(car)
```

Loading required package: carData

```
leveneTest(variable ~ group, data = data)
```

Warning in leveneTest.default(y = y, group = group, ...): group coerced to factor.

Levene's Test for Homogeneity of Variance (center = median)

| | Df | F value | Pr(>F) |
|-------|----|---------|--------|
| group | 1 | 0.0769 | 0.7885 |
| | 8 | | |

36 Chapter 13: Non-Parametric Alternatives

36.1 Wilcoxon Signed Rank Test

```
wilcox.test(before, after, paired = TRUE)
```

```
Warning in wilcox.test.default(before, after, paired = TRUE): cannot compute  
exact p-value with ties
```

Wilcoxon signed rank test with continuity correction

```
data: before and after  
V = 15, p-value = 0.05676  
alternative hypothesis: true location shift is not equal to 0
```

36.2 Mann-Whitney U Test

```
wilcox.test(group1, group2)
```

```
Warning in wilcox.test.default(group1, group2): cannot compute exact p-value  
with ties
```

Wilcoxon rank sum test with continuity correction

```
data: group1 and group2  
W = 24.5, p-value = 0.01597  
alternative hypothesis: true location shift is not equal to 0
```

36.3 Kruskal-Wallis Test

Non-parametric alternative to ANOVA.

```
# Sample data frame for Kruskal-Wallis Test
data <- data.frame(
  score = c(85, 88, 90, 80, 82, 85, 78, 80, 83, 90, 92, 95, 88, 90, 91),
  group = rep(c("A", "B", "C"), times = 5)
)
kruskal.test(score ~ group, data = data)
```

Kruskal-Wallis rank sum test

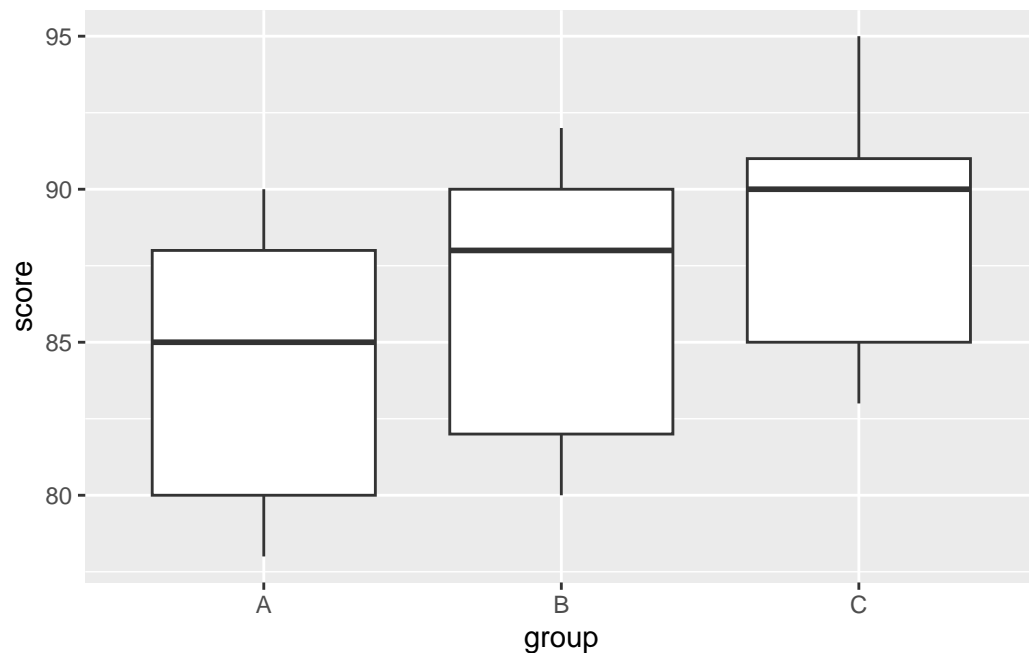
data: score by group

Kruskal-Wallis chi-squared = 2.2329, df = 2, p-value = 0.3274

37 Chapter 14: Visualizing Statistical Results

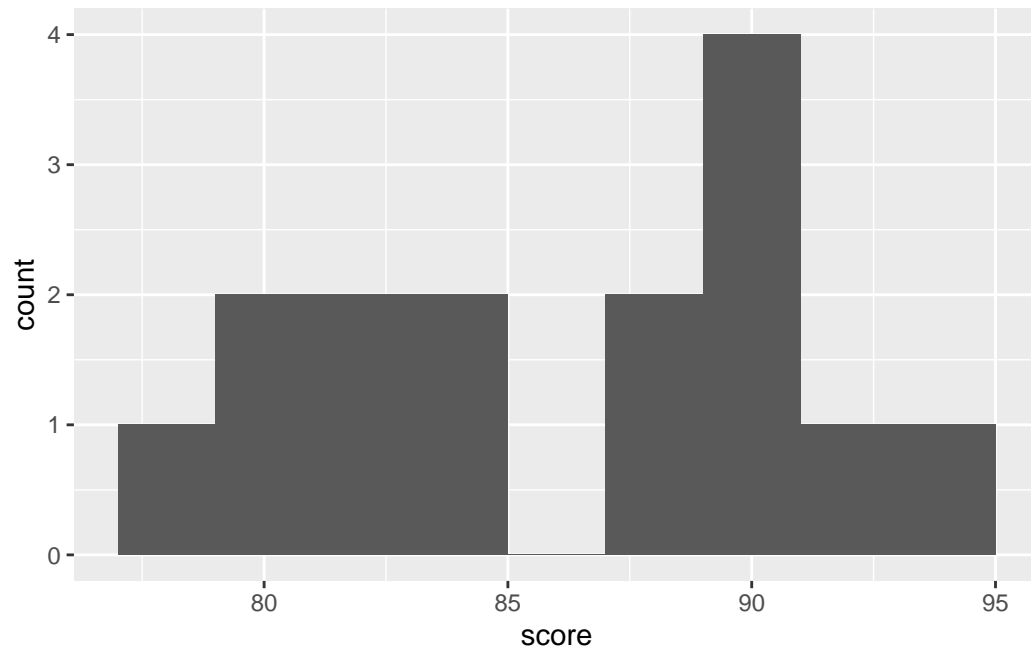
37.1 Boxplots

```
ggplot(data, aes(x = group, y = score)) +  
  geom_boxplot()
```



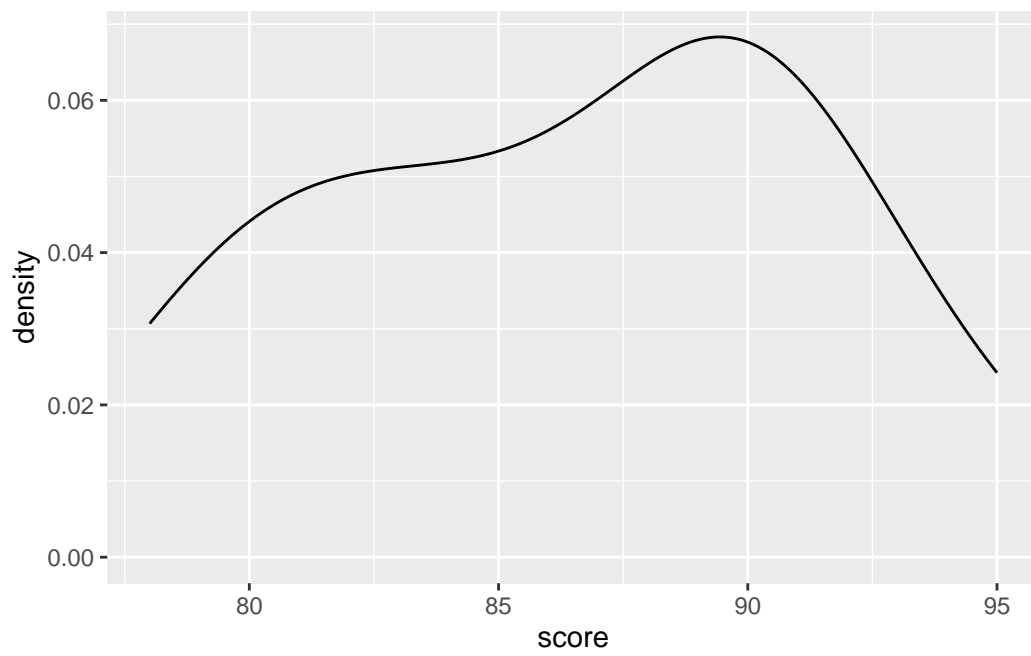
37.2 Histograms

```
ggplot(data, aes(x = score)) +  
  geom_histogram(binwidth = 2)
```



37.3 Density Plot

```
ggplot(data, aes(x = score)) +  
  geom_density()
```



38 Chapter 15: RKWard (GUI-R) Tips

- Use menu-based analysis for beginners
- Save and export plots easily
- Integrate with R scripts for reproducibility

38.1 Summary: Basic Statistics using GUI-R (RKWard)

This eBook, authored by Dr. Harsh Pradhan (Assistant Professor at the Institute of Management Studies, Banaras Hindu University), serves as a comprehensive guide to understanding and applying basic statistical concepts, particularly in the GUI-based software RKWard (GUI-R).

Key Highlights: 1. Descriptive Statistics Covers measures of central tendency (mean, median, mode) and variability (range, variance, standard deviation). Introduces standard error and its role in estimating population parameters. 2. Inferential Statistics Introduces the Central Limit Theorem and how it forms the foundation for many statistical techniques. Confidence intervals are explained both theoretically and with practical calculations. 3. T-Tests (Student's t) Explains one-sample, independent-sample, and paired-sample t-tests. Includes step-by-step computation and GUI-R implementation. Includes interpretation of p-values, degrees of freedom, and test statistics. 4. Analysis of Variance (ANOVA) Covers one-way, two-way, and repeated measures ANOVA. Focuses on the F-statistic, assumptions, and post-hoc analyses. Discusses partitioning of variance into systematic and unsystematic components. 5. Effect Size and Statistical Power Introduces Cohen's d, eta-squared, and power analysis. Emphasizes that statistical significance does not always imply practical importance. 6. Assumption Testing Tests for normality (Shapiro-Wilk, QQ plot). Tests for homogeneity of variance (Levene's test). Highlights when to use non-parametric alternatives. 7. Non-Parametric Tests Introduces Wilcoxon signed-rank, Mann-Whitney U, and Kruskal-Wallis tests as robust alternatives to parametric methods. 8. Data Visualization in R Demonstrates use of boxplots, histograms, and density plots using ggplot2. Provides example R code for reproducibility. 9. GUI-R (RKWard) Usage Offers practical steps for using GUI-R for all statistical techniques covered. Designed to bridge the gap for learners unfamiliar with command-line R.