

Visual Model Summary

Parmeshvar

2025-07-06

Table of contents

1	Introduction	8
1.1	Lecture notes	8
1.2	Moodle website	8
2	Week 1	12
2.1	Module 1: Introduction to Statistics	12
2.1.1	Pre-Requisites	12
2.1.2	Agenda	12
2.1.3	Meaning of Statistics	12
2.1.4	Nature of Statistics	13
2.1.5	Uses of Statistics	13
2.1.6	Limitations of Statistics	13
2.2	Misuse of Statistics	13
2.3	Module 2: Mathematics vs Statistics	14
2.4	Module 3: Software-Based Statistical Revolution	14
2.4.1	Popular Statistical Software	14
2.4.2	GUI vs CLI	14
2.4.3	Recommended GUI Tools for R	15
2.4.4	Installing RKWard on Ubuntu	15
2.5	Module 4: Understanding Variables	15
2.5.1	What is a Variable?	15
2.5.2	R Definition:	15
3	Week 2	22
3.1	Introduction	22
3.1.1	Purpose of the eBook	22
3.1.2	Who Should Read This?	22
3.1.3	What You'll Learn	22
3.2	1. Fundamentals of Statistics	22
3.2.1	1.1 What is Statistics?	22
3.2.2	1.2 Key Objectives	23
3.2.3	1.3 Types of Statistics	23
3.3	2. Types of Data	23
3.3.1	2.1 Classification of Data	23
3.4	3. Descriptive Statistics	24
3.4.1	3.1 Measures of Central Tendency	24
3.4.2	3.2 The Mean	24
3.4.3	3.3 The Median	25
3.4.4	3.4 The Mode	25
3.4.5	3.5 Comparison Table	26

3.5	4. Measures of Variability	26
3.5.1	4.1 Why Measure Variability?	26
3.5.2	4.2 Range	26
3.5.3	4.3 Quartiles and Interquartile Range	27
3.5.4	4.4 Variance	27
3.5.5	4.5 Standard Deviation	28
3.5.6	4.6 Coefficient of Variation (CV)	28
3.5.7	4.7 Moment-Based Measures	28
3.6	5. Probability Fundamentals	29
3.6.1	5.1 Introduction to Probability	29
3.6.2	5.2 Key Definitions	29
3.6.3	5.3 Types of Events	29
3.6.4	5.4 Classical Probability	29
3.6.5	5.5 Probability Rules	29
3.6.6	5.6 Conditional Probability	30
3.7	6. Discrete Probability Distributions	31
3.7.1	6.1 Bernoulli Distribution	31
3.7.2	6.2 Binomial Distribution	31
3.8	7. Continuous Distributions	32
3.8.1	7.1 Normal Distribution	32
3.8.2	7.2 Standard Normal Distribution	32
3.9	8. Visualizing Data	33
3.9.1	8.1 Frequency Distribution	33
3.9.2	8.2 Histogram	33
3.9.3	8.3 Boxplot (Box-and-Whisker Plot)	33
3.9.4	8.4 Scatter Plot	33
3.10	9. Practical Applications	34
3.10.1	9.1 Business Use Cases	34
3.10.2	9.2 Education and Research	34
3.11	10. Using RKWard	34
3.11.1	10.1 What is RKWard?	34
3.11.2	10.2 Installation Guide	34
3.11.3	10.3 Sample RKWard Activities	34
3.11.4	10.4 Using R Code in RKWard	35
4	Week 3	36
4.1	Introduction	36
4.1.1	Importance of Statistics	36
4.1.2	Overview of Topics	36
4.2	Understanding Populations and Samples	37
4.2.1	Population	37
4.2.2	Sample	37
4.2.3	Why Use Samples?	37
4.2.4	Relation Between Population & Sample	37
4.3	Hypotheses and Errors	37
4.3.1	Hypothesis Defined	37
4.3.2	Types of Errors	38
4.3.3	Significance Level ()	38

4.4	Inferential Statistics	38
4.4.1	Purpose	38
4.4.2	Common Techniques	38
4.4.3	Sampling Techniques	38
4.4.4	Central Limit Theorem (CLT)	39
4.5	Descriptive Statistics	39
4.5.1	Measures of Central Tendency	39
4.5.2	Measures of Dispersion	40
4.5.3	Measures of Shape	40
4.6	Graphical Methods	40
4.6.1	Histogram	40
4.6.2	R Data Types and Structures	41
4.6.3	Comparing R vs Excel vs GUI-R (RKWard)	41
4.6.4	Installing RKWard (Ubuntu)	41
4.6.5	Teaching Tools in RKWard	41
4.6.6	GUI-Based Statistical Tools	42
4.7	Linear Regression in R	42
4.7.1	What is Linear Regression?	42
4.7.2	Code Example	42
4.8	Fit model	42
4.9	Summary	42
4.9.1	Adjusted R-squared	43
4.9.2	Normal Distribution	43
4.9.3	Data Import Techniques	43
4.9.4	Working with the RKWard Interface	43
4.9.5	Spreadsheet Concepts	43
4.9.6	Advantages	43
4.9.7	Limitations	44
4.9.8	Advanced Plots and Techniques	44
4.9.9	Common R Packages for Statistics	46
4.9.10	Introduction to Command Line	46
4.10	Windows Terminal	46
4.10.1	Git + R Project Example	46
4.10.2	Fallacies and Bias: Real-World Cautions	47
4.10.3	Future Applications of Statistics	47
4.10.4	Practice Challenges	47
4.10.5	Key Takeaways	47
5	Week 4	48
5.1	Introduction	48
5.2	Course Overview	48
5.2.1	Course Name	48
5.2.2	Instructor Profile	48
5.2.3	Learning Objectives	48
5.3	Chapter 1: Fundamental Concepts	48
5.3.1	Descriptive Statistics	48
5.3.2	Standard Error	49
5.3.3	Central Limit Theorem	49

5.3.4	Confidence Intervals	50
5.4	Chapter 2: Estimation	50
5.4.1	Types of Estimates	50
5.4.2	Parameter vs Statistic	50
5.5	Chapter 3: Hypothesis Testing	50
5.6	Chapter 4: Student's T-Test	50
5.6.1	Types	50
5.6.2	One-Sample T-Test Example	51
5.6.3	Test Statistic	51
5.6.4	Degrees of Freedom	51
5.6.5	Decision Rule	51
5.6.6	T-Test in GUI-R	51
5.7	Chapter 5: ANOVA	52
5.7.1	Purpose	52
5.7.2	Post-Hoc Tests	52
5.7.3	ANOVA in GUI-R	53
5.8	Chapter 6: GUI-R Workflow	53
5.9	Chapter 7: Advanced Concepts	53
5.9.1	Variance Partitioning	53
5.9.2	Degrees of Freedom	53
5.9.3	Chi-Square and F Distribution	53
5.9.4	Univariate, Bivariate, Multivariate	54
5.9.5	Parametric Test Assumptions	54
5.9.6	Effect Size	54
5.9.7	Power of a Test	54
5.10	Conclusion	54
5.11	References	55
5.12	Chapter 8: Advanced T-Test Applications	55
5.12.1	Paired Sample T-Test	55
5.12.2	Independent Samples T-Test	55
5.12.3	One-Sample T-Test with GUI-R	56
5.13	Chapter 9: More on Confidence Intervals	56
5.13.1	Visualizing Confidence Intervals in R	56
5.14	Chapter 10: Robust ANOVA Models	57
5.14.1	Two-Way ANOVA	57
5.14.2	Repeated Measures ANOVA	58
5.15	Chapter 11: Effect Size Measures	58
5.15.1	Cohen's d	58
5.15.2	Eta-Squared (η^2)	59
5.16	Chapter 12: Statistical Assumptions Checking	59
5.16.1	Normality	59
5.16.2	Homogeneity of Variance	60
5.17	Chapter 13: Non-Parametric Alternatives	61
5.17.1	Wilcoxon Signed Rank Test	61
5.17.2	Mann-Whitney U Test	61
5.17.3	Kruskal-Wallis Test	61
5.18	Visualizing Statistical Results	62

5.19	Boxplots	62
5.19.1	Histograms	62
5.19.2	Density Plot	63
5.20	RKward (GUI-R) Tips	64
5.20.1	Summary: Basic Statistics using GUI-R (RKward)	64
6	Week 5	65
6.1	1. Introduction	65
6.2	2. Lecture 24 – Deep Dive: Correlation	65
6.2.1	2.1 What is Correlation?	65
6.2.2	2.2 Types of Correlation and Use Cases	65
6.2.3	2.3 Pearson, Spearman, Kendall Comparison	66
6.3	Add non-linear data	66
6.4	Pearson (linear)	66
6.5	Spearman (rank, monotonic)	66
6.6	Kendall (ordinal)	66
6.7	install.packages(“ggm”)	67
6.8	Simulate data	67
6.9	Generate all pairwise correlations	67
6.10	Visualize matrix with corrplot	67
6.11	3. Lecture 25 – One-Way ANOVA (Detailed)	68
6.11.1	3.1 Concept Overview	68
6.11.2	3.2 ANOVA Table Example	68
6.11.3	3.3 R Code – One-Way ANOVA	68
6.12	Homogeneity check	69
6.13	6. Lecture 28 – Simple Linear Regression	70
6.13.1	6.1 Theory Refresher	70
6.13.2	6.2 Example in R	70
6.13.3	install.packages(“scatterplot3d”)	71
6.14	10. Lecture 32 – Logistic Regression	71
6.14.1	10.1 When to Use	71
6.14.2	10.2 Logistic Function	72
6.14.3	10.3 R Example: Predicting Admission	72
7	Week 7	74
7.1	1. Introduction	74
7.2	2. Time Series Analysis	74
7.2.1	2.1 Overview of Time Series Data	74
7.3	Load and visualize example data	74
7.4	Simulate joint probability	75
7.5	Prior probabilities	75
7.6	Bayes’ formula	75
7.7	4. Expected Value and Bivariate Variables	76
7.7.1	4.1 Expected Value Basics	76
8	Week 8	81
8.1	1. Introduction	81

8.2	2. Effect Size and Cohen's d	81
8.2.1	Interpretation of d:	81
8.2.2	R Code Example (Cohen's d)	81
8.3	Load required package	81
8.4	Load your data (CSV format)	81
8.5	Independent groups Cohen's d	82
8.6	One-sample mean vs population mean	82
8.7	4. Using flexplot: Examples and Best Practices	82
8.7.1	4.1 Univariate Visualization	82
8.8	Visualizing continuous DV vs categorical IV	83
8.9	Convert pass/fail variable to factor	83
8.10	Logistic visualization	83
8.11	What's Next in Part 3?	85
8.12	10. Simulation: Effect Size and Visual Inference	85
8.12.1	10.1 Simulate Cohen's d with Flexplot	85
8.13	13. Model Summary with Visual + Numeric Layers	86

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, 2010

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

[Orcid ID](#)

[Google Scholar](#)

[Youtube ID](#)

[Academic Profile](#)

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

1.1 Lecture notes

Download from [here](#).

1.2 Moodle website

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

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
Week 2	1	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
	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

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
	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
	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

Week	Main Topic	Subtopic	Video	PDF Resource
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
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

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: eyJyZXNvdXJjZURpciI6Ii4iLCJib29rSXRlbVR5cGUiOiJjaGFwdGVyIiwiaYm9va01`

## 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

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

4.1 Introduction

4.1.1 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**.

4.1.2 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
-

4.2 Understanding Populations and Samples

4.2.1 Population

The complete set of all units of interest. Examples:

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

4.2.2 Sample

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

4.2.3 Why Use Samples?

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

4.2.4 Relation Between Population & Sample

Population → Sample → Statistic → Inference → Population Parameter

4.3 Hypotheses and Errors

4.3.1 Hypothesis Defined

A hypothesis is a testable assumption about a population.

4.3.1.1 Null Hypothesis (H_0)

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

4.3.1.2 Alternative Hypothesis (H_A)

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

4.3.2 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)

4.3.3 Significance Level ()

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

4.4 Inferential Statistics

4.4.1 Purpose

- Estimate unknown population parameters
- Test hypotheses
- Predict outcomes

4.4.2 Common Techniques

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

4.4.3 Sampling Techniques

4.4.3.1 1. Simple Random Sampling

Every unit has equal probability.

4.4.3.2 2. Systematic Sampling

Pick every k th element.

4.4.3.3 3. Stratified Sampling

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

4.4.3.4 4. Cluster Sampling

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

4.4.4 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:

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

4.5 Descriptive Statistics

4.5.1 Measures of Central Tendency

4.5.1.1 Mean

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

4.5.1.2 Median

Middle value in an ordered dataset.

4.5.1.3 Mode

Most frequent value.

4.5.2 Measures of Dispersion

4.5.2.1 Range

$$Range = Max - Min$$

4.5.2.2 Variance

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

4.5.2.3 Standard Deviation

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

4.5.3 Measures of Shape

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

4.6 Graphical Methods

4.6.1 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.

4.6.2 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

4.6.3 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

4.6.4 Installing RKWard (Ubuntu)

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

4.6.5 Teaching Tools in RKWard

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

4.6.6 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

4.7 Linear Regression in R

4.7.1 What is Linear Regression?

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

4.7.1.1 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

4.7.2 Code Example

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

4.8 Fit model

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

4.9 Summary

```
summary(model)
```

4.9.1 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).

4.9.2 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

4.9.3 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")`

4.9.4 Working with the RKWard Interface

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

4.9.5 Spreadsheet Concepts

Structure

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

4.9.6 Advantages

Easy data entry Visual inspection Good for small datasets

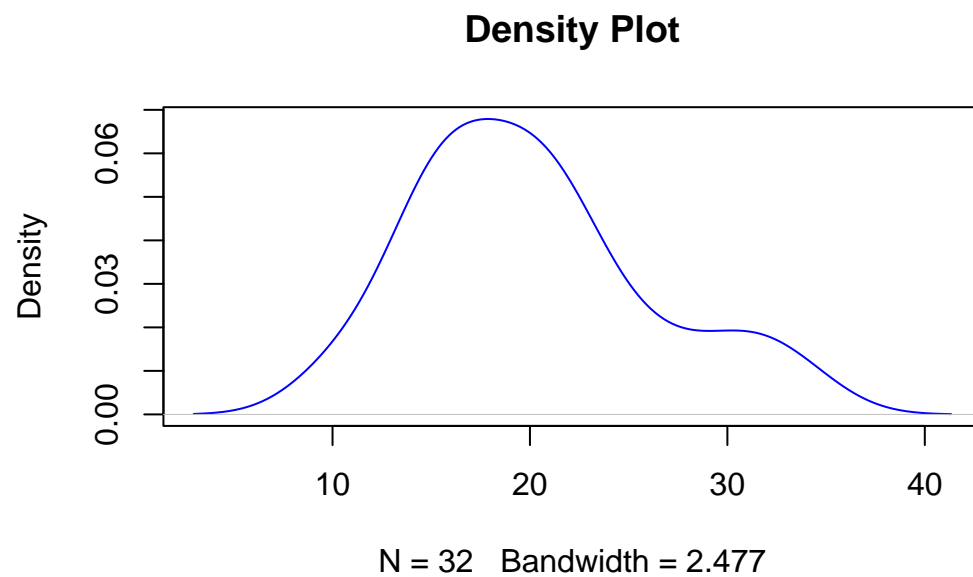
4.9.7 Limitations

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

4.9.8 Advanced Plots and Techniques

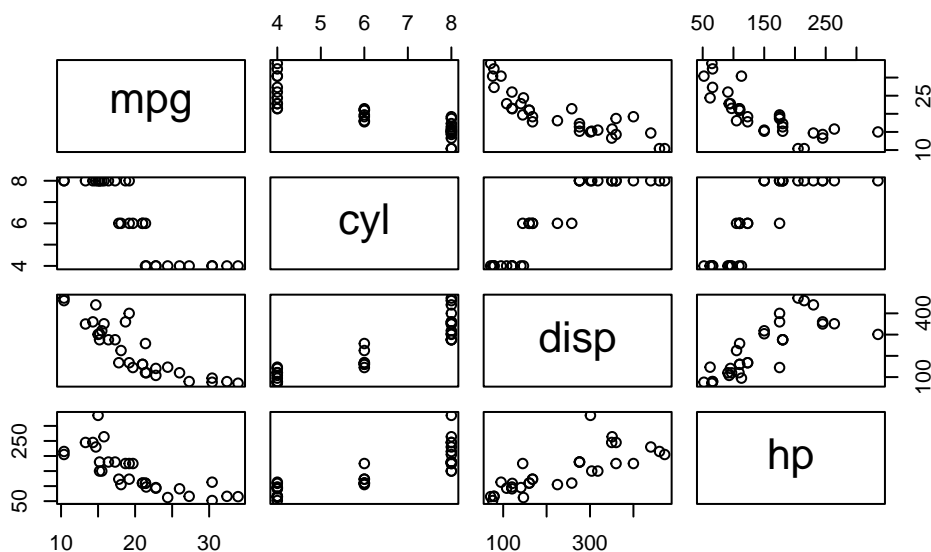
Density Plot

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



Pair Plot

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



Correlation Matrix

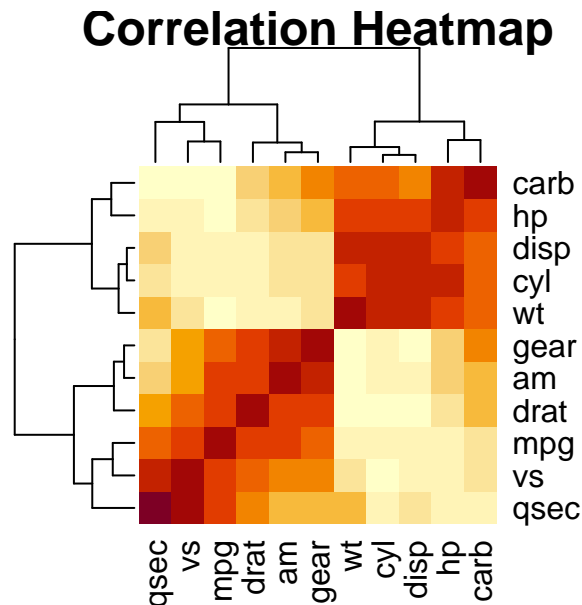
```
cor(mtcars)
```

	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	1.0000000	0.7940588	0.05753435
gear	-0.21268223	0.2060233	0.79405876	1.0000000	0.27407284
carb	-0.65624923	-0.5696071	0.05753435	0.2740728	1.0000000

Heatmap

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



4.9.9 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

4.9.10 Introduction to Command Line

4.10 Windows Terminal

```
cd..mkdirmy_projectdir
```

Linux Terminal

```
cd mkdirstats_projectls -l
```

4.10.1 Git + R Project Example

```
gitinitgitclonehttps://github.com/username/project.git
```

4.10.2 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

4.10.3 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

4.10.4 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

4.10.5 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.

5 Week 4

5.1 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.

5.2 Course Overview

5.2.1 Course Name

Basic Statistics using GUI-R (RKWard)

5.2.2 Instructor Profile

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

[Faculty Profile](#)

5.2.3 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

5.3 Chapter 1: Fundamental Concepts

5.3.1 Descriptive Statistics

5.3.1.1 Central Tendency

- Mean
- Median

- Mode

5.3.1.2 Dispersion

- Range
- Variance
- Standard Deviation

5.3.1.3 Example:

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

```
[1] 5.2
```

```
median(data)
```

```
[1] 5
```

```
sd(data)
```

```
[1] 1.923538
```

5.3.2 Standard Error

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

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

5.3.3 Central Limit Theorem

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

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

5.3.4 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.

5.4 Chapter 2: Estimation

5.4.1 Types of Estimates

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

5.4.2 Parameter vs Statistic

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

5.5 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

5.6 Chapter 4: Student's T-Test

5.6.1 Types

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

5.6.2 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
```

5.6.3 Test Statistic

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

5.6.4 Degrees of Freedom

$$df = n - 1$$

5.6.5 Decision Rule

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

5.6.6 T-Test in GUI-R

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

5.7 Chapter 5: ANOVA

5.7.1 Purpose

Used when comparing means across 3+ groups.

5.7.1.1 One-Way ANOVA Formula

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

Where:

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

5.7.1.2 Assumptions

- Normality
- Homogeneity of variance
- Independence

5.7.1.3 Example Table

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

5.7.2 Post-Hoc Tests

Run if ANOVA is significant to locate pairwise differences.

5.7.3 ANOVA in GUI-R

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

5.8 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

5.9 Chapter 7: Advanced Concepts

5.9.1 Variance Partitioning

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

Explained Terms	Unexplained Terms
Systematic	Random
Predictive	Error
Deterministic	Noise

5.9.2 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.

5.9.3 Chi-Square and F Distribution

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

5.9.4 Univariate, Bivariate, Multivariate

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

5.9.5 Parametric Test Assumptions

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

If assumptions violated → use non-parametric test.

5.9.6 Effect Size

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

Used for comparison across studies.

5.9.7 Power of a Test

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

Higher power → lower chance of Type II error

Power increases with sample size, effect size

5.10 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.

5.11 References

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

5.12 Chapter 8: Advanced T-Test Applications

5.12.1 Paired Sample T-Test

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

5.12.1.1 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
```

5.12.2 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
```

5.12.3 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

5.13 Chapter 9: More on Confidence Intervals

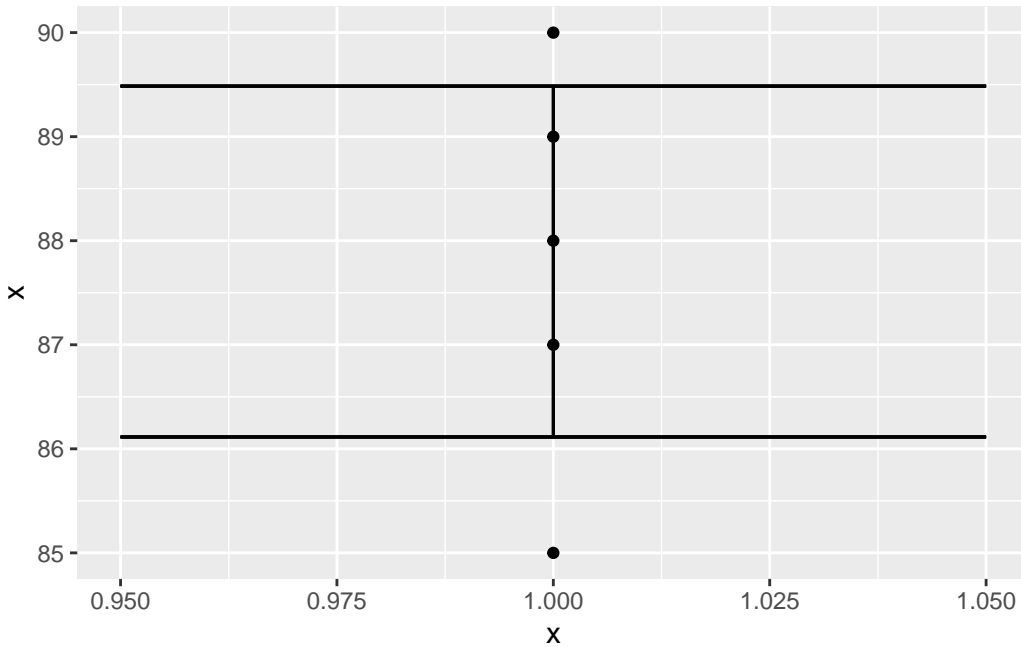
5.13.1 Visualizing Confidence Intervals in R

```
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)
```

5.14 Chapter 10: Robust ANOVA Models

5.14.1 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

5.14.2 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 *
```

5.15 Chapter 11: Effect Size Measures

5.15.1 Cohen's d

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

Where s_p is the pooled standard deviation.

5.15.1.1 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
```

5.15.2 Eta-Squared (η^2)

Used for ANOVA:

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

5.16 Chapter 12: Statistical Assumptions Checking

5.16.1 Normality

Use Shapiro-Wilk test:

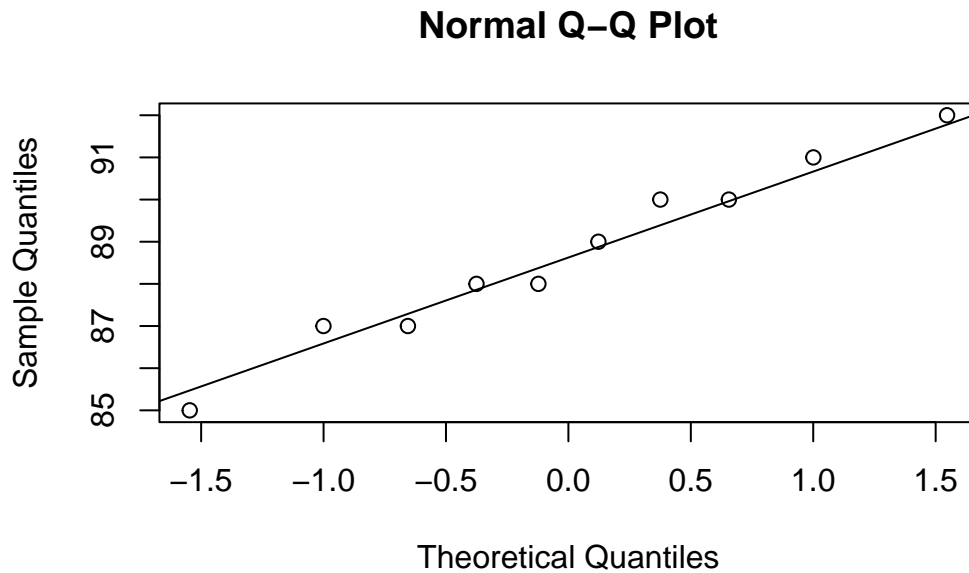
```
# Sample data frame for normality test
data <- data.frame(variable = c(88, 90, 85, 87, 89, 91, 92, 88, 90, 87))
shapiro.test(data$variable)
```

Shapiro-Wilk normality test

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

Visualize:

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



5.16.2 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
```

5.17 Chapter 13: Non-Parametric Alternatives

5.17.1 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

5.17.2 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

5.17.3 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

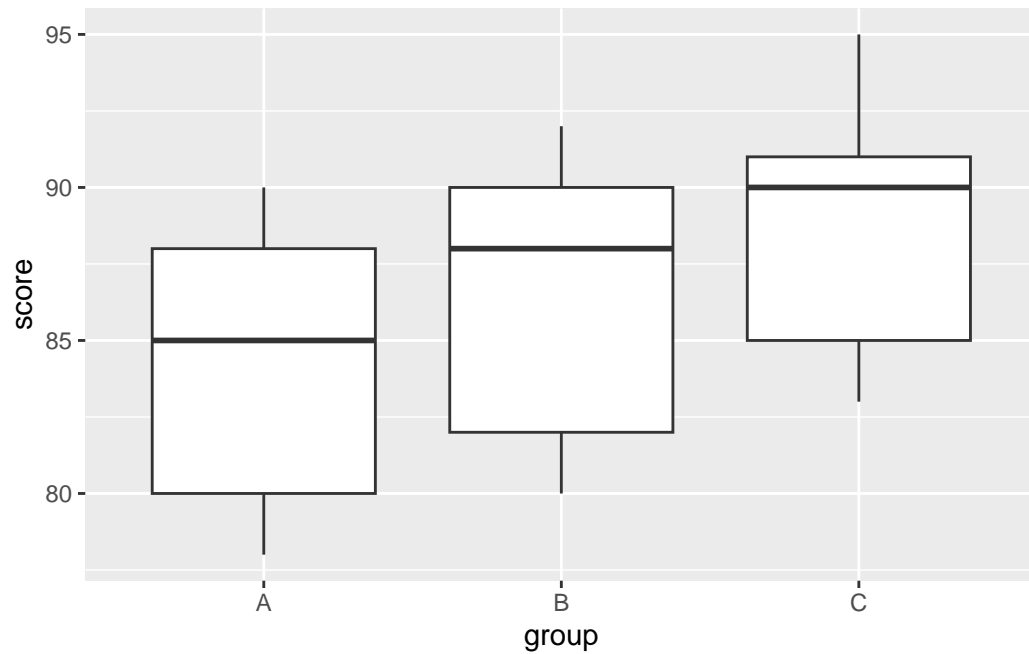
data: score by group

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

5.18 Visualizing Statistical Results

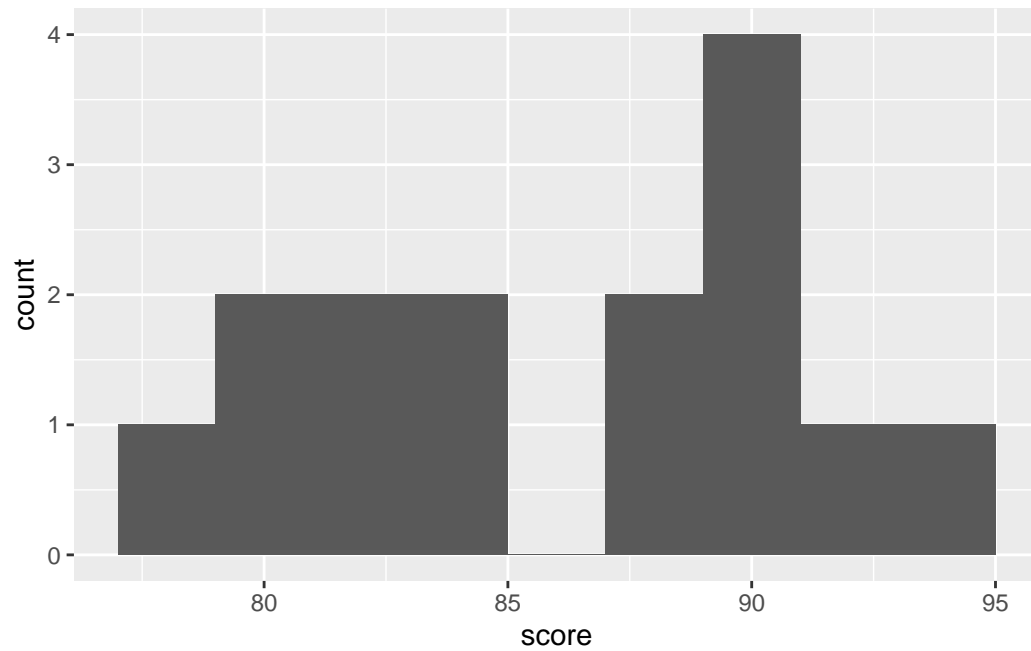
5.19 Boxplots

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



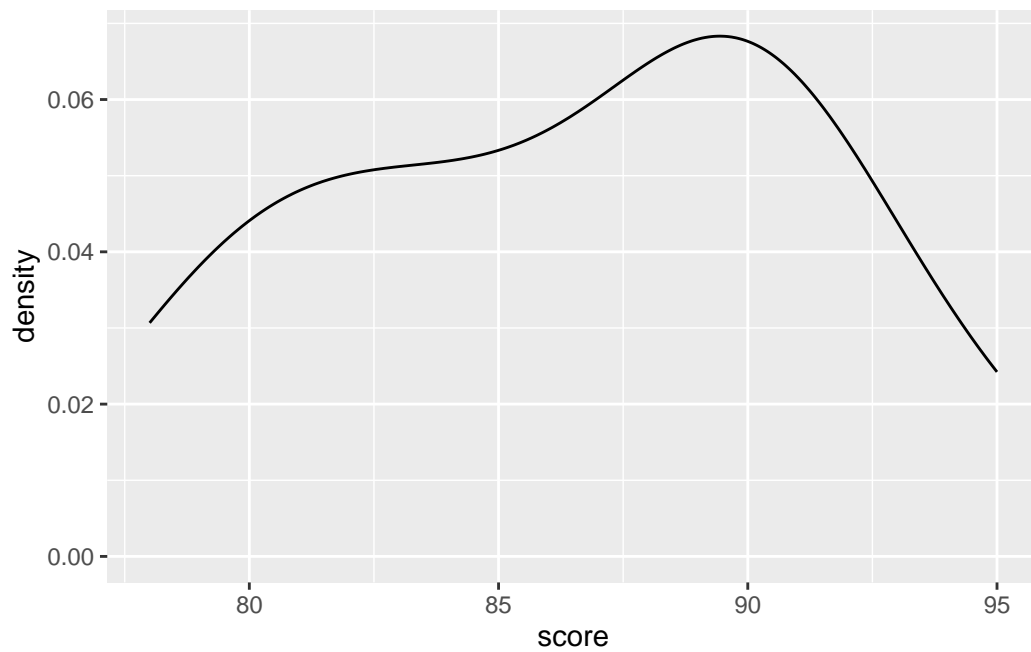
5.19.1 Histograms

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



5.19.2 Density Plot

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



5.20 RKWard (GUI-R) Tips

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

5.20.1 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.

6 Week 5

6.1 1. Introduction

Welcome to **Week 5** of *Basic Statistics using GUI-R (RKWard)*, where we cover **relationship testing, correlation, regression, ANOVA**, and related diagnostics in **depth**, using both theory and R-based implementation.

This book is designed to:

- Clarify statistical concepts visually
 - Use real data simulations
 - Empower you with reproducible R/RKWard workflows
 - Prepare you to run statistical diagnostics and build interpretations
-

6.2 2. Lecture 24 – Deep Dive: Correlation

6.2.1 2.1 What is Correlation?

Correlation is a statistical measure that expresses the **extent to which two variables are linearly related**.

6.2.1.1 Theory

- If variable X increases as Y increases → **Positive correlation**
- If variable X increases as Y decreases → **Negative correlation**
- If there's no linear trend → **Zero correlation**

Pearson's r ranges from -1 to +1.

6.2.2 2.2 Types of Correlation and Use Cases

Data Type	Correlation Type	Use Case Example
Nominal	Phi	Gender vs. Yes/No Preferences
Dichotomous	Point-Biserial	Pass/Fail vs. Exam Score
Ordinal/Rank	Spearman/Kendall	Rank in class vs. Test anxiety
Ratio/Interval	Pearson	Height vs. Weight
Multivariate	Partial Correl.	Control confounders

6.2.3 2.3 Pearson, Spearman, Kendall Comparison

```
{r} ## Simulate linear data set.seed(123) x <- rnorm(100) y <- 2 * x + rnorm(100)
```

6.3 Add non-linear data

```
z <- x^2 + rnorm(100)
```

6.4 Pearson (linear)

```
cor(x, y, method = "pearson")
```

6.5 Spearman (rank, monotonic)

```
cor(x, z, method = "spearman")
```

6.6 Kendall (ordinal)

```
cor(x, z, method = "kendall") 2.4 Visualizing Correlations ## Visualization library(ggplot2) data <- data.frame(x, y, z)
```

```
ggplot(data, aes(x = x, y = y)) + geom_point() + geom_smooth(method = "lm", se = FALSE, color = "blue") + labs(title = "Scatter Plot with Linear Fit", x = "X", y = "Y")
```

```
ggplot(data, aes(x = x, y = z)) + geom_point(color = "darkred") + labs(title = "Non-Linear Relationship", x = "X", y = "Z") 2.5 Correlation Matrix in RKWard Steps:
```

Load data into RKWard.

Navigate to Statistics → Summaries → Correlation Matrix.

Choose the appropriate variables.

Choose correlation type (Pearson, Spearman).

Run and interpret the matrix output.

2.6 Partial Correlation in R When you want to compute the correlation between two variables while controlling for a third:

6.7 `install.packages("ggm")`

```
library(ggm) X1 <- rnorm(100) X2 <- X1 + rnorm(100, sd = 0.5) X3 <- rnorm(100) pcor(c("X1",  
"X2", "X3"), cov(cbind(X1, X2, X3)))
```

 Interpretation: This tells you the pure correlation between X1 and X2, controlling for X3.

2.7 R Code to Automate All

6.8 Simulate data

```
set.seed(100) data <- data.frame( A = rnorm(100), B = rnorm(100), C = rnorm(100) )
```

6.9 Generate all pairwise correlations

```
cor(data)
```

6.10 Visualize matrix with `corrplot`

```
library(corrplot) corrplot(cor(data), method = "color", tl.col = "black", addCoef.col = "black")
```

 2.8 Spearman vs Pearson – When to Use? Use Pearson when data is normally distributed, continuous, and linear.

Use Spearman when data is ordinal, ranked, or non-linear but monotonic.

Kendall's Tau is more robust for small sample sizes.

Next Up: Part 2/4 will include:

- One-Way ANOVA full theory + math
- Repeated Measures ANOVA (detailed)
- Visualization of F-distributions
- MANOVA + N-Way examples
- 10+ R code exercises

6.11 3. Lecture 25 – One-Way ANOVA (Detailed)

6.11.1 3.1 Concept Overview

Analysis of Variance (ANOVA) is used when comparing the means of three or more groups.

6.11.1.1 Formula Breakdown

- SSM (Sum of Squares Model): Variation between groups
- SSR (Sum of Squares Residual): Variation within groups
- SST (Total): Total variation

F-Ratio:

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SSM/df_M}{SSR/df_R}$$

6.11.2 3.2 ANOVA Table Example

Source	SS	df	MS	F
Between	461.64	3	153.88	8.27
Within	167.42	9	18.60	
Total	629.08	12		

6.11.3 3.3 R Code – One-Way ANOVA

```
group1 <- c(28, 36, 38, 31) group2 <- c(32, 33, 40) group3 <- c(47, 43, 52) group4 <- c(40, 47, 45)
```

```
score <- c(group1, group2, group3, group4) group <- factor(rep(c("Hunter", "Farming", "Natural", "Industrial"), times=c(4,3,3,3)))
```

```
data <- data.frame(score, group) anova_model <- aov(score ~ group, data=data) summary(anova_model)
```

3.4 Post-Hoc Analysis (Tukey HSD)

TukeyHSD(anova_model)

3.5 Visualize Group Differences

```
boxplot(score ~ group, data = data, col = c("lightblue", "pink", "lightgreen", "yellow"))
```

4. Lecture 26 – Repeated Measures ANOVA

4.1 Theory Repeated measures involve the same subjects measured under multiple conditions.

Aspect Repeated Measures Between-Subjects Subjects Same across treatments Different per group

Variability Control Higher (less noise) Lower Efficiency More efficient Requires more samples

4.2 R Code – Repeated Measures

```
library(ez) subject <- factor(rep(1:10, each=3)) treatment <- factor(rep(c("Pre", "Mid", "Post"),
times=10)) score <- c(rnorm(10, 65), rnorm(10, 70), rnorm(10, 75)) rm_df <- data.frame(subject,
treatment, score)
```

```
ezANOVA(data=rm_df, dv=score, wid=subject, within=treatment)
```

```
library(ggplot2) ggplot(rm_df, aes(x=treatment, y=score, group=subject, color=subject)) +
geom_line() + geom_point() + theme_minimal() + labs(title="Repeated Measures ANOVA
Plot")
```

5. Lecture 27 – MANOVA and N-Way ANOVA 5.1 What is MANOVA? Multivariate Analysis of Variance (MANOVA) extends ANOVA to multiple dependent variables.

Example Use Case:

Investigating how teaching methods affect:

Exam scores

Class participation

Homework submission

5.2 R Code – MANOVA

```
y1 <- rnorm(30, 60, 5) y2 <- rnorm(30, 70, 6) y3 <- rnorm(30, 80, 4) method <- factor(rep(c("A",
"B", "C"), each=10))
```

```
manova_model <- manova(cbind(y1, y2, y3) ~ method) summary(manova_model)
```

```
df <- expand.grid( Teaching = c("Traditional", "Interactive"), Gender = c("Male", "Female"), Rep
= 1:20 ) df$Score <- rnorm(80, mean = 70, sd = 5)
```

```
model_nway <- aov(Score ~ Teaching * Gender, data = df) summary(model_nway)
```

```
Plot {r} interaction.plot(df$Teaching, df$Gender, df$Score, col=c("red", "blue"))
```

5.4 Interaction Plot {r} interaction.plot(df\$Teaching, df\$Gender, df\$Score, col=c("red", "blue")) 5.5 Assumptions of ANOVA Assumption Check Method Tool Normality QQ Plot, Shapiro Test shapiro.test() Homogeneity Levene's/Bartlett's Test car::leveneTest() Independence Design-level assurance Design phase

5.6 Assumption Check in R {r} # Normality check shapiro.test(residuals(anova_model))

6.12 Homogeneity check

```
library(car) leveneTest(score ~ group, data = data)
```

```
curve(df(x, df1=3, df2=9), from=0, to=10, col="blue", lwd=2, ylab="Density", main="F-
distribution df(3,9)") abline(v=8.27, col="red", lwd=2, lty=2) legend("topright", legend=c("F =
8.27"), col="red", lty=2)
```

```
set.seed(2024) group_A <- rnorm(10, mean=50) group_B <- rnorm(10, mean=51) group_C <-
rnorm(10, mean=50.5)
```

```
score <- c(group_A, group_B, group_C) group <- factor(rep(c("A", "B", "C"), each=10))
```

```
df <- data.frame(score, group) aov_model <- aov(score ~ group, data=df) summary(aov_model)
```

End of Part 2/4. Part 3 includes Regression (Simple, Multiple, Non-linear), VIF, Residuals, and Advanced Modeling

6.13 6. Lecture 28 – Simple Linear Regression

6.13.1 6.1 Theory Refresher

Linear regression predicts a **dependent variable (Y)** using an **independent variable (X)**.

Model Equation:

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

Where:

- β_0 = Intercept
 - β_1 = Slope
 - ϵ = Error term
-

6.13.2 6.2 Example in R

```
study_time <- c(2, 3, 4, 5, 6) grades <- c(50, 60, 65, 70, 75)
```

```
model <- lm(grades ~ study_time) summary(model)
```

6.3 Regression Line Visualization

```
plot(study_time, grades, main="Simple Regression", xlab="Study Time", ylab="Grades")  
abline(model, col="blue", lwd=2)
```

6.4 Interpret Coefficients

coef(model) Intercept: Grade when study time = 0

Slope: Grade increases per hour of study

6.5 Residual Plots

```
par(mfrow=c(2,2)) plot(model)
```

Top-left: Residuals vs Fitted

Bottom-left: Scale-Location

Top-right: QQ Plot

Bottom-right: Residuals vs Leverage

6.6 Confidence Intervals

```
confint(model)
```

7. Lecture 29 – Multiple Regression 7.1 Add More Predictors

```
df <- data.frame( Exam = c(50, 55, 60, 65, 70), Hours = c(2, 3, 4, 5, 6), Sleep = c(7, 6.5, 6, 5.5, 5)
) multi_model <- lm(Exam ~ Hours + Sleep, data = df) summary(multi_model) 7.2 Check VIF
(Multicollinearity)
```

```
library(car) vif(multi_model) VIF > 5 → multicollinearity warning VIF > 10 → serious problem
```

7.3 Partial Residual Plots

```
avPlots(multi_model) 7.4 Plot 3D Regression Plane
```

6.13.3 install.packages("scatterplot3d")

```
library(scatterplot3d) scatterplot3d(dfHours, dfSleep, df$Exam, highlight.3d=TRUE, type="h",
angle=55, color="darkgreen", pch=16) 8. Lecture 30 – Polynomial and Non-Linear Regression 8.1
Simulating Non-linear Relationship
```

```
x <- seq(0, 10, 0.1) y <- 5 + 2 * x^2 + rnorm(length(x), 0, 5) plot(x, y, main="Non-linear Pattern",
pch=19) 8.2 Polynomial Regression
```

```
poly_model <- lm(y ~ poly(x, 2)) summary(poly_model)
```

```
lines(x, predict(poly_model), col="blue", lwd=2) 8.3 Compare with Linear Fit
```

```
linear_model <- lm(y ~ x) lines(x, predict(linear_model), col="red", lwd=2, lty=2) leg-
end("topleft", legend=c("Poly", "Linear"), col=c("blue", "red"), lty=c(1,2)) 8.4 Residual
Analysis
```

```
par(mfrow=c(1,2)) plot(poly_model$fitted.values, poly_model$residuals, main="Polynomial Residu-
als") plot(linear_model$fitted.values, linear_model$residuals, main="Linear Residuals") 8.5 Curve
Fitting with nls()
```

```
x <- seq(0, 10, length.out=100) y <- 2 * exp(0.3 * x) + rnorm(100, sd=3)
```

```
nls_model <- nls(y ~ a * exp(b * x), start=list(a=2, b=0.3)) summary(nls_model)
```

```
lines(x, predict(nls_model), col="purple", lwd=2) 9. Lecture 31 – Model Evaluation Metrics 9.1
R2 and Adjusted R2
```

```
summary(multi_model)r.squaredsummary(multi_model)adj.r.squared 9.2 MSE and RMSE
```

```
pred <- predict(multi_model) actual <- df$Exam residuals <- actual - pred mse <-
mean(residuals^2) rmse <- sqrt(mse)
```

6.14 10. Lecture 32 – Logistic Regression

6.14.1 10.1 When to Use

Logistic regression is used when the **dependent variable is categorical** (typically binary: 0/1, Yes/No, Pass/Fail).

6.14.2 10.2 Logistic Function

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

6.14.3 10.3 R Example: Predicting Admission

```
df <- data.frame( Admit = c(1,1,0,1,0,0,1,1,0,0), Score = c(80,85,60,90,55,40,88,83,59,52) )
```

```
logit_model <- glm(Admit ~ Score, data=df, family="binomial") summary(logit_model) 10.4  
Probability Prediction
```

```
df$Prob <- predict(logit_model, type="response") df 10.5 ROC Curve
```

```
library(pROC) roc_obj <- roc(dfAdmit, dfProb) plot(roc_obj, col="darkgreen") auc(roc_obj)  
10.6 Classification Table
```

```
dfPred <- ifelse(dfProb > 0.5, 1, 0) table(Predicted = dfPred, Actual = dfAdmit) 11. Lec-  
ture 33 – Chi-Square Test 11.1 Categorical Independence Used when evaluating if two categorical  
variables are independent.
```

11.2 Example: Gender vs Department Choice

```
gender <- c("Male", "Male", "Female", "Female") dept <- c("Science", "Arts", "Science", "Arts")  
counts <- c(30, 20, 25, 25)
```

```
chi_df <- data.frame(Gender=rep(gender, counts), Dept=rep(dept, counts)) tbl <- table(chi_dfGender, chi_dfDept)  
chisq.test(tbl) 12. Lecture 34 – Non-Parametric Tests 12.1 When to Use Data is not normally  
distributed
```

Ordinal data or small sample sizes

12.2 Mann–Whitney U

```
group1 <- c(45, 50, 60, 55) group2 <- c(70, 75, 80, 85) wilcox.test(group1, group2) 12.3 Kruskal–  
Wallis (Non-parametric ANOVA)
```

```
g1 <- c(10, 20, 30) g2 <- c(40, 50, 60) g3 <- c(70, 80, 90) kw_df <- data.frame( score = c(g1,  
g2, g3), group = factor(rep(c("A", "B", "C"), each=3)) ) kruskal.test(score ~ group, data=kw_df)  
12.4 Wilcoxon Signed-Rank
```

```
before <- c(60, 70, 65, 80) after <- c(62, 75, 68, 82) wilcox.test(before, after, paired=TRUE) 13.  
Case Study – Social Media & Mental Health 13.1 Dataset Simulation
```

```
set.seed(100) n <- 100 hours <- rnorm(n, 3, 1.5) stress <- 10 + 1.2 * hours + rnorm(n)
```

```
df <- data.frame(hours, stress) model <- lm(stress ~ hours, data=df) summary(model) 13.2 Vi-  
sual
```

```
plot(hours, stress, main="Social Media Use vs Stress", pch=19) abline(model, col="red", lwd=2)  
13.3 Interpretation Positive slope → More hours = more stress
```

R² tells how well hours predict stress

14. 50 Multiple Choice Questions (MCQs) Q1. Pearson's r is best used when: Data is ordinal
Data is continuous and normally distributed
Data has outliers
You want to rank variables
Q2. Which test compares more than 2 independent means? t-test
ANOVA
Chi-Square
Correlation
Q3. A VIF of 12 means: No multicollinearity
Severe multicollinearity
Perfect fit
Homoscedasticity
15. Exercises Exercise 1: One-Way ANOVA on Fake Marketing Data Generate three ad strategies and test which gives highest customer conversions.
Exercise 2: Correlate temperature and ice cream sales Include scatterplot, Pearson's r , regression line.
Exercise 3: Logistic regression predicting credit approval Predict using income and debt ratio.
Exercise 4: Chi-Square on survey data Test independence of satisfaction vs. purchase intention.
Exercise 5: Repeated Measures ANOVA Simulate 10 people tested across 3 time points.
16. Glossary Term Definition ANOVA Test for differences in means across groups Regression Predict numerical output from inputs Correlation Measure of linear association between two variables R^2 Proportion of variance explained by model AIC Akaike Information Criterion – model quality metric VIF Variance Inflation Factor – checks multicollinearity Logistic Regression Used for binary outcome prediction Chi-Square Test for independence between two categorical variables
17. Appendix 17.1 RKWard Menus Correlation → Statistics → Summaries → Correlation Matrix
ANOVA → Analysis → ANOVA → One-Way or Repeated
Plots → Graphics → Histogram / Boxplot / Scatterplot
Regression → Analysis → Linear Models
- 17.2 Troubleshooting Issue Solution “object not found” Check variable names (case-sensitive)
Plot doesn't show Use `print(plot_name)` or run outside R chunk Model output blank Use `summary(model)` instead of just `model` Package not found Install using `install.packages("name")`

7 Week 7

7.1 1. Introduction

This eBook focuses on key statistical topics covered in **Week 7** of the course *Basic Statistics using GUI-R (RKWard)*. From **time series forecasting** to **Bayesian probability** and **discrete distributions**, each concept is explored with R-based demonstrations, code implementations, and visual outputs.

7.2 2. Time Series Analysis

7.2.1 2.1 Overview of Time Series Data

7.3 Load and visualize example data

```
install.packages("TSA") library(TSA) data(tempdub) plot(tempdub, main="Monthly Temperature in Dubuque")
```

Trend: Long-term increase or decrease

Seasonality: Predictable recurring patterns

Cyclic: Irregular, long-term fluctuations

2.2 Data Import and Price Fetching

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

```
first.date <- Sys.Date() - 90 last.date <- Sys.Date() stocks <- c("TCS.NS") tcs_prices <- BatchGetSymbols(tickers = stocks, first.date, last.date) write.csv(tcs_prices$data, "tcs.csv")
```

2.3 Handling Seasonality

```
rt <- diff(log(tempdub), 12) # Seasonal difference for monthly data plot(rt, main = "Seasonally Differenced Series")
```

```
library(tseries) adf.test(rt) # Test for stationarity
```

Monthly Dummies

```
month <- season(tempdub) m1 <- lm(tempdub ~ month - 1) summary(m1) resid <- residuals(m1) adf.test(resid)
```

2.4 Trend Extraction & Detrending

```
sim <- rnorm(100, mean = 0, sd = 10) x <- 5 + time(sim)*3 + ts(sim) x <- ts(x) plot(x)
```

```
model2 <- lm(x ~ time(x)) resid2 <- resid(model2) adf.test(resid2)
```

2.5 Smoothing Techniques

Simple Moving Average (SMA)

```
library(forecast) ts_data <- ts(c(10, 15, 20, 25, 30, 35, 40)) sma <- ma(ts_data, order = 3)
plot(sma, type = 'l', col = 'blue') Exponential Moving Average (EMA)
```

```
library(TTR) data <- c(23, 45, 67, 34, 56, 78, 90) ts_data <- ts(data) ema <- EMA(ts_data, n
= 3) plot(ema, type = 'l', col = 'darkgreen') 2.6 Forecasting Models Naive Forecasting: Future =
last value
```

ARIMA:

```
library(forecast) fit <- auto.arima(AirPassengers) forecast(fit, h = 12) plot(forecast(fit, h = 12))
ETS Models:
```

```
ets_model <- ets(AirPassengers) plot(forecast(ets_model)) 2.7 Accuracy Metrics
```

```
actual <- c(100, 110, 120) pred <- c(98, 112, 119)
```

```
MAE <- mean(abs(actual - pred)) RMSE <- sqrt(mean((actual - pred)^2)) MAPE <-
mean(abs((actual - pred)/actual)) * 100
```

```
print(c(MAE = MAE, RMSE = RMSE, MAPE = MAPE)) 3. Conditional Probability & Bayes'
Theorem 3.1 Conditional Probability If  $P(B) > 0$ , then:
```

7.4 Simulate joint probability

```
joint <- matrix(c(0.1, 0.2, 0.2, 0.5), nrow = 2) P_A_given_B <- joint[1,2] / (joint[1,2] + joint[2,2])
print(P_A_given_B) 3.2 Bayes' Theorem
```

7.5 Prior probabilities

```
P_user <- 0.05 P_pos_given_user <- 0.9 P_neg_given_nonuser <- 0.8 P_nonuser <- 1 - P_user
P_pos_given_nonuser <- 1 - P_neg_given_nonuser
```

7.6 Bayes' formula

```
P_user_given_pos <- (P_pos_given_user * P_user) / ((P_pos_given_user * P_user) +
(P_pos_given_nonuser * P_nonuser))
```

```
print(P_user_given_pos) 3.3 Real-Life Applications Medical Testing
```

Spam Filtering

Credit Risk Modeling

7.7 4. Expected Value and Bivariate Variables

7.7.1 4.1 Expected Value Basics

For discrete variable X :

$$E(X) = \sum x_i \cdot P(x_i)$$

```
x <- c(1, 2, 3, 4)
p <- c(0.1, 0.3, 0.4, 0.2)
expected_value <- sum(x * p)
print(expected_value)
```

4.2 Linearity of Expectation

If $Y = aX + b$:

```
a <- 3
b <- 5
E_X <- expected_value
E_Y <- a * E_X + b
print(E_Y)
```

4.3 Bivariate Distributions

Example: Coin Toss (from PPT)

Let:

X = number of heads

Y = |heads - tails|

Then, for 3 coin tosses:

```
joint_pmf <- matrix(c(
  0, 0, 0, 1/8,
  0, 3/8, 0, 0,
  0, 3/8, 0, 0,
  0, 0, 0, 1/8
), nrow = 4, byrow = TRUE)
```

```
colnames(joint_pmf) <- c("Y=0", "Y=1", "Y=2", "Y=3")
rownames(joint_pmf) <- c("X=0", "X=1", "X=2", "X=3")
print(joint_pmf)
```

4.4 Marginal Probabilities

```
# Marginal P(X)
rowSums(joint_pmf)
```

```
# Marginal P(Y)
colSums(joint_pmf)
```

5. Discrete Distributions

5.1 Hypergeometric Distribution

```
get_probability <- function(N, K, n, k) {  
  choose(K, k) * choose(N - K, n - k) / choose(N, n)  
}  
  
N <- 10  
K <- 6  
n <- 5  
possible_k <- 0:n  
  
probabilities <- sapply(possible_k, function(k) get_probability(N, K, n, k))  
  
barplot(probabilities, names.arg = possible_k,  
        xlab = "White Balls in Sample", ylab = "Probability",  
        col = "lightblue", main = "Hypergeometric Distribution")
```

5.2 Poisson Distribution

```
lambda <- 2  
values <- 0:10  
prob_pois <- dpois(values, lambda)  
  
barplot(prob_pois, names.arg = values,  
        main = "Poisson(=2)", col = "orange")
```

5.3 Negative Binomial Distribution

```
p <- 0.70  
r <- 5  
attempts <- 5:20  
pmf <- dnbinom(attempts - r, size = r, prob = p)  
  
plot(attempts, pmf, type = "h", lwd = 2, col = "blue",  
     main = "Negative Binomial Distribution",  
     xlab = "Attempts", ylab = "Probability")
```

5.4 Geometric Distribution

```
p <- 0.3  
x_vals <- 1:20  
geo_prob <- dgeom(x_vals - 1, prob = p)  
  
plot(x_vals, geo_prob, type = "h", col = "darkgreen",  
     main = "Geometric Distribution",  
     xlab = "Trial", ylab = "P(success at k-th trial)")
```

6. Practical Applications

6.1 Bayesian Inference in R

Bayes inference example with normal prior/posterior:

```
library(ggplot2)

prior <- rnorm(10000, mean = 0.3, sd = 0.1)
likelihood <- rnorm(10000, mean = 0.35, sd = 0.05)
posterior <- (prior + likelihood)/2

df <- data.frame(
  value = c(prior, likelihood, posterior),
  dist = factor(rep(c("Prior", "Likelihood", "Posterior"), each = 10000))
)

ggplot(df, aes(x = value, fill = dist)) +
  geom_density(alpha = 0.5) +
  labs(title = "Bayesian Updating")
```

6.2 Forecasting in Finance and Healthcare
 Finance: Time series of stock returns

Healthcare: Spread of diseases

```
# Example of forecast in time series
library(forecast)
fit <- auto.arima(AirPassengers)
forecast_vals <- forecast(fit, h = 24)
plot(forecast_vals, main = "AirPassengers Forecast")
```

6.3 Effect Size Estimation

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

cohensD(c(3.2, 3.4, 3.5), mu = 3.0)
Effect size values:
```

Small: 0.2

Medium: 0.5

Large: 0.8

This section wraps up with:

ARIMA modeling

Stationarity & Unit Root tests (ADF)

Residual analysis

Advanced diagnostics

Summary & references

7. Advanced Statistical Concepts

7.1 Stationarity and Unit Root Testing

A **stationary time series** has constant mean and variance over time. Its essential for:

- Forecasting
- Valid modeling
- Avoiding spurious regression

Unit Root: Augmented Dickey-Fuller (ADF) Test

```
library(tseries)
set.seed(42)
x <- cumsum(rnorm(100)) # non-stationary random walk
plot.ts(x, main = "Simulated Random Walk")
```

```
adf.test(x) # Likely non-stationary (p > 0.05)
```

7.2 Detrending Time Series

```
t <- time(x)
trend_model <- lm(x ~ t)
resid_trend <- resid(trend_model)
plot(resid_trend, main = "Detrended Series")
adf.test(resid_trend) # Residuals should now be stationary
```

7.3 ARIMA Modeling

Autoregressive Integrated Moving Average

AR(p): Autoregression

I(d): Differencing

MA(q): Moving average

```
library(forecast)
auto.arima(AirPassengers)
Full Workflow
```

```
tsdata <- AirPassengers
plot(tsdata)
```

```
# Step 1: Stationarity check
adf.test(tsdata) # May need differencing
```

```
# Step 2: Model Selection
```

```
fit <- auto.arima(tsdata)
summary(fit)
```

```
# Step 3: Forecasting
fc <- forecast(fit, h = 12)
plot(fc)
```

7.4 ACF & PACF Plots

Used for identifying model orders:

```
acf(diff(log(AirPassengers)))
pacf(diff(log(AirPassengers)))
```

7.5 Residual Diagnostics

```
checkresiduals(fit) # From forecast package
Box.test(residuals(fit), lag = 20, type = "Ljung-Box")
```

7.6 Forecast Accuracy

```
actuals <- window(AirPassengers, start = c(1960,1))
preds <- forecast(fit, h = 12)$mean
accuracy(preds, actuals)
```

8. Summary

This eBook covered advanced Week 7 content with practical R implementation:

Topic Key Concepts & Tools

Time Series Analysis TSA, decomposition, ADF test, ARIMA

Conditional Probability Bayes theorem, real-life problems

Expected Value Joint PMFs, linearity of expectation

Discrete Distributions Poisson, Hypergeometric, Negative Binomial

Forecasting Techniques SMA, EMA, ETS, ARIMA

Bayesian Applications Posterior inference, medical testing

Model Evaluation AIC, BIC, RMSE, MAPE, residuals

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

```
`{=html}
```

```
<!-- quarto-file-metadata: eyJyZXNvdXJjZURpciI6Ii4iLCJib29rSXRlbVR5cGUiOiJjaGFwdGVyIiwiaYm9va0l
```


8 Week 8

8.1 1. Introduction

This module explores the powerful integration of visual analytics and statistical reasoning. While traditional models often rely on tabular outputs, the **flexplot** package and similar tools highlight the importance of **graphical modeling**, especially in response to the **replication crisis**. The week also emphasizes how GUIs like **RKward** and **RStudio** serve different user bases for statistical analysis.

8.2 2. Effect Size and Cohen's d

Effect size quantifies the **magnitude** of the difference, independent of sample size. One of the most common effect size measures is **Cohen's d**, which compares two means.

8.2.1 Interpretation of d:

d	Meaning
0.2	Small effect
0.5	Medium effect
0.8	Large effect

8.2.2 R Code Example (Cohen's d)

8.3 Load required package

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

8.4 Load your data (CSV format)

```
my.csv.data <- read.csv("yourdata.csv")
```

8.5 Independent groups Cohen's d

```
lsr::cohensD(my.csv.data[["CSE_1"]], my.csv.data[["CSE_2"]])
```

8.6 One-sample mean vs population mean

```
lsr::cohensD(my.csv.data[["CSE_1"]], mu = 3.9)
```

 Practical Use: Effect size helps you understand practical significance, especially in behavioral research where p-values alone are insufficient.

Note: Effect sizes should always accompany inferential statistics to avoid overreliance on significance testing.

3. Understanding flexplot: Graphical Statistical Modeling The flexplot package allows for intuitive, formula-driven visual modeling. It uses GLM-style formulas like $y \sim x_1 + x_2$, bringing clarity between statistical models and their graphical representations.

Key Features: Visualize univariate, bivariate, and multivariate models

Supports linear, logistic, and mixed models

Matches graphical output directly with statistical models

Requires only one line of code for most use cases

Installation and Setup:

```
install.packages("flexplot") library(flexplot) library(cowplot) # For arranging multiple plots
```

 More Coming in Part 2: Univariate & Bivariate flexplot() demos

GLM integration

Paneling, Ghost Lines, Beeswarm Visuals

Overlap handling and jitter control

8.7 4. Using flexplot: Examples and Best Practices

8.7.1 4.1 Univariate Visualization

```
flexplot(CSE_1 ~ 1, data = my.csv.data)
```

 Plots raw data (jittered) with mean overlay

Useful for outlier detection and distribution shape

4.2 Bivariate Continuous vs Categorical

8.8 Visualizing continuous DV vs categorical IV

`flexplot(CSE_1 ~ Gender, data = my.csv.data)` Automatically creates beeswarm or violin plots

Overlay: mean \pm error bars

Jitter is used to prevent overlap of points

4.3 Continuous DV vs Continuous IV

`flexplot(CSE_1 ~ Age, data = my.csv.data)` Shows scatterplot + best-fit line

Adds error ribbons

Outliers stand out visually

4.4 Multiple Predictors (Additive Models)

`flexplot(CSE_1 ~ Age + Gender, data = my.csv.data)` Panels by Gender

Linear fits across Age

Helps uncover interaction

4.5 Logistic Regression Visualization

8.9 Convert pass/fail variable to factor

`my.csv.data$Pass <- as.factor(my.csv.data$Pass)`

8.10 Logistic visualization

`flexplot(Pass ~ Hours, data = my.csv.data, family = "binomial")` 5. RKWard vs RStudio: Interface & Functionality Feature RKWard RStudio Target Users Beginners, GUI-centric Coders, devs, advanced users Data Handling Spreadsheet-like Tidyverse-friendly Plots Auto-generated via dialogs `ggplot2` required manually Statistical Models GUI for t-tests, ANOVA Syntax for all models

Conclusion: RKWard is ideal for non-programmers, while RStudio is better for reproducible analysis via code and markdown.

6. Cognitive Fit and Visual Communication Flexplot builds upon Cognitive Fit Theory — visual representations should match the task and viewer's expectation.

Key Graph Types in flexplot Type Best For Beeswarm Small-to-medium samples Violin Density + mean overlay Ghost Lines Slope visualization across panels Panels 2–3 categorical moderators

7. Advanced Flexplot Controls 7.1 Ghost Lines for Slope Tracking

`flexplot(mpg ~ wt + cyl, data = mtcars)` Panels by cyl

Gray reference slope: overall trend

Colored slope: panel-specific

7.2 Model Overlays

`flexplot(CSE_1 ~ CSE_2 + Gender, data = my.csv.data)` Adds regression lines

Includes model summaries in plot captions

7.3 Added Plot (Influence Visualization)

`model <- lm(CSE_1 ~ CSE_2 + Age, data = my.csv.data)` `added.plot(model)` Visualizes the unique contribution of predictors

Residual scatter by regressor

8. Association, AVPs, and Repeated Measures 8.1 Visualizing Correlation

`flexplot(mpg ~ hp, data = mtcars)` Adds correlation line

Includes Pearson's r

8.2 Repeated Measures (Paneling)

`flexplot(score ~ time + condition, data = repeated_df)` Each condition as panel

Time as predictor

Fits separate lines

8.3 Binned Paneling (Continuous Moderators)

`flexplot(CSE_1 ~ Age + Income, data = my.csv.data)` Age: X-axis

Income: Panel bins (equal-width)

Visualizes moderation effects

8.4 Jitter, Transparency, Point Customization

`flexplot(CSE_1 ~ Age + Gender, data = my.csv.data, jitter = 0.3, alpha = 0.5, point.size = 2)` 9.

Interactive Plots and R Markdown Integration

`install.packages("plotly")` `library(plotly)` `p <- flexplot(mpg ~ wt + cyl, data = mtcars)` `ggplotly(p)`
Adds interactivity Quarto Embedding markdown

`flexplot(CSE_1 ~ Gender, data = my.csv.data)`

8.11 What's Next in Part 3?

- Full-scale simulation for effect size
- Reproducible workflows
- Custom function design
- Summary + export instructions

This final section includes:

Simulation for effect size

Custom model visuals

Reproducible workflows

Summary + rendering/export notes

8.12 10. Simulation: Effect Size and Visual Inference

8.12.1 10.1 Simulate Cohen's d with Flexplot

```
set.seed(123) group1 <- rnorm(50, mean = 5, sd = 1) group2 <- rnorm(50, mean = 6.2, sd = 1)
group <- factor(rep(c("A", "B"), each = 50)) score <- c(group1, group2)
```

```
sim_df <- data.frame(group, score)
```

```
library(lsr) cohensD(score ~ group, data = sim_df) # Should return d 1.2
```

```
flexplot(score ~ group, data = sim_df) 10.2 Power and Confidence Visualization
```

```
library(pwr) pwr.t.test(d = 0.8, power = 0.8, sig.level = 0.05, type = "two.sample") 10.3 Monte Carlo Effect Size Estimation
```

```
sim_d <- replicate(1000, { g1 <- rnorm(30, 5, 1) g2 <- rnorm(30, 6, 1) cohensD(g1, g2) })
```

```
hist(sim_d, breaks = 50, col = "lightblue", main = "Simulated Cohen's d Distribution") abline(v = mean(sim_d), col = "red") 11. Visual Inference in Teaching Overlay Raw + Model Together
```

```
flexplot(mpg ~ wt + cyl, data = mtcars) Cyl = Panel
```

Gray slope = overall

Color slope = per panel

R² and p-values appear below

12. Workflow: Reproducible Visual Analytics in R 12.1 Data Import

```
df <- read.csv("CSE_scores.csv") str(df) 12.2 Visualization Plan Start with flexplot()
```

Panel by categorical moderators

Add continuous predictors

Use `added.plot()` to show incremental effect

Report both visualization + model summary

12.3 R Markdown Report markdown

```
library(flexplot) flexplot(score ~ gender + age, data = df)
```

8.13 13. Model Summary with Visual + Numeric Layers

```
model <- lm(CSE_1 ~ CSE_2 + Age + Gender, data = my.csv.data) summary(model)
```

```
added.plot(model) # Visual version of unique effect 14. Combining flexplot with ggplot2
```

```
p1 <- flexplot(CSE_1 ~ CSE_2 + Gender, data = my.csv.data) p2 <- ggplot(my.csv.data,  
aes(CSE_2, CSE_1)) + geom_point() + geom_smooth(method = "lm")
```

```
cowplot::plot_grid(p1, p2, labels = c("Flexplot", "GGplot")) 15. Summary Concept Tool Used  
Effect Size cohensD() from lsr Graphical Modeling flexplot() Simulation Monte Carlo Association  
Plot Slope Panels Influence Plot added.plot() Interactive Graphs ggplotly()
```