The Importance of Summary Statistics and Techniques for Creating Them in R

Baharalsadat Bahadori - 400897377 - bahadori.baharalsadat@stud.hs-fresenius.de

Introduction

- Summary statistics are the **first step** in understanding any dataset.
- They help us see:
  - Typical values (mean, median)
  - How data is spread (variance, SD)
  - Patterns and anomalies
  - Possible errors or outliers
- In this presentation, I will explain:
  - Why summary statistics are important
  - The types of summary statistics
  - Techniques to compute them in R

Why Summary Statistics Matter

- Summary statistics help us **quickly understand** large datasets.
- They show:
  - Central values (mean, median)
  - Spread of data (range, variance, standard deviation)
  - Distribution shape (skewness, kurtosis)
- They help detect:
  - Outliers
  - Incorrect values
  - Missing data
  - Data entry mistakes
- They make analysis **faster**, **simpler**, and **more accurate**.

Types of Summary Statistics

1. Measures of Central Tendency

- **Mean** – average value
- **Median** – middle value
- **Mode** – most frequent value

2. Measures of Spread

- **Range** – difference between max and min
- **Variance** – how far values are from the mean
- **Standard Deviation** – average spread of data

3. Measures of Shape

- **Skewness** – direction of tilt
- **Kurtosis** – how peaked or flat the distribution is

4. Count & Frequency

- **Count of observations**
- **Frequency tables**

1. Measures of Central Tendency

Mean, Median, Mode

Example:

Student	Score
80 90 85 90	90 75 85 90

mean = 84 (average performance)

2. Measures of Spread

Range, Variance, Standard Deviation

Example:

Employee	Salary
1 2 3	2000 2100 4000

range = 2000 to 4000

Techniques to Compute Summary Statistics in R

Base R: Simple Summary Statistics

Example Dataset

Heights of 5 students:

```r
heights <- c(160, 165, 170, 155, 180)
```

Base R Summary Statistics

```r
mean(heights)      # average height
median(heights)    # middle value
sd(heights)        # standard deviation
summary(heights)   # full summary
```

dplyr Summary Statistics

```r
library(dplyr)

df <- data.frame(heights)

df %>%
  summarise(
    mean_height = mean(heights, na.rm = TRUE),
    median_height = median(heights, na.rm = TRUE),
    sd_height = sd(heights, na.rm = TRUE)
  )
```

Explanation

- `summarise()` creates new summary values.
- Each line inside summarise calculates one statistic.
- `na.rm = TRUE` removes missing values before calculation.

Grouped Summary Statistics

```r
library(dplyr)

# Example dataset
df_grouped <- data.frame(
  group = c("A", "A", "B", "B", "B"),
  value = c(10, 12, 8, 15, 14)
)

# Grouped summary
df_grouped %>%
  group_by(group) %>%
  summarise(
    avg_value = mean(value),
    count     = n()
  )
```

Explanation

- `group_by(group)` splits the dataset into categories (A and B).
- `summarise()` calculates summary statistics *inside each group*.
- This helps compare groups easily (e.g., group A vs group B).

🎯 Student Exercise: Try It Yourself

Dataset

Scores: **45, 60, 75, 80, 90**

Your Tasks

- Calculate the **mean** of the scores
- Find the **median**
- Calculate the **standard deviation**
- Use `summary(scores)` and observe the output
- Bonus: Convert the scores into a data frame and print it

Questions to Discuss

- What does the mean tell you about these scores?
- Are the values close together or widely spread?
- Does the summary output match your manual calculations?

✅ Solution to the Exercise

Dataset

Scores: **45, 60, 75, 80, 90**

Calculations

- **Mean:** 70
- **Median:** 75
- **Standard Deviation:** 17.08 (approx)
- **Summary Output Includes:**
  - Min = 45
  - 1st Quartile = 60
  - Median = 75
  - Mean = 70
  - 3rd Quartile = 80
  - Max = 90

📝 Conclusion

- Summary statistics provide a **quick understanding** of any dataset.
- They help identify typical values, variation, and possible errors.
- **Base R**, **dplyr**, and **grouped summaries** offer easy ways to compute them.
- Summary statistics are essential before visualization or modeling.
- With just a few lines of R code, we can extract **clear, meaningful insights** from data.