Unit III: Descriptive Statistics in R

Jamuna S Murthy
Assistant Professor
Department of CSE

Academic Year - 2025-26

1 Introduction

Descriptive statistics summarize and describe the main features of a dataset. R provides several built-in functions and packages to compute these measures. These statistics are essential to get a quick understanding of the data before further analysis.

2 Measures of Central Tendency

These measures describe the center of the data distribution: mean, median, and mode.

2.1 Mean

The mean is the average of all data points.

```
data1 <- c(10, 20, 30, 40, 50)
mean(data1)  # Output: 30
```

2.2 Median

The median is the middle value in an ordered dataset.

```
data2 <- c(12, 15, 14, 13, 16)
median(data2)  # Output: 14
```

2.3 Mode (Custom Function)

R does not have a built-in mode function. We can define one:

```
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
data3 <- c(2, 4, 4, 4, 5, 6, 6)
get_mode(data3)  # Output: 4
```

- `unique(v)`: Extracts the unique values from the vector `v`.
- `match(v, uniqv)`: Returns the positions of each element in `v` as they match `uniqv`.
- `tabulate(...)`: Counts how many times each position appears, creating a frequency table.
- `which.max(...)`: Finds the index of the highest frequency.
- `uniqv[...]`: Returns the actual value from the unique list corresponding to the mode.

3 Measures of Variability

These measures describe how much the data varies or spreads.

3.1 Range

The range is the difference between the maximum and minimum values.

```
data4 <- c(5, 10, 15, 20, 25)
range(data4)  # Output: 5 25
```

3.2 Variance

The variance measures the average squared deviation from the mean.

```
var(data4)  # Output: 62.5
```

3.3 Standard Deviation

The standard deviation is the square root of the variance.

```
sd(data4)  # Output: 7.9 (approximately)
```

4 Skewness and Kurtosis

To compute skewness and kurtosis, we use the `moments` package.

4.1 Installing and Loading the Package

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

4.2 Skewness

Skewness tells us about the asymmetry of the distribution.

```
data5 <- c(10, 20, 30, 40, 100)
skewness(data5)  # Output: positive skew
```

4.3 Kurtosis

Kurtosis tells us about the peakedness or flatness of a distribution.

```
kurtosis(data5)  # Output: > 3 means leptokurtic (sharp peak)
```

5 Full Example: Summary of a Dataset

```
data <- c(5, 10, 15, 20, 25, 30, 100)
summary(data)  # Provides min, 1st Qu., median, mean, 3rd Qu., max

mean(data)      # Average
median(data)    # Middle value
sd(data)        # Standard deviation
range(data)     # Minimum and maximum
skewness(data)  # Asymmetry
kurtosis(data)  # Tailedness
```

6 Dataset Examples

6.1 Example 1: Student Marks Analysis

```
marks <- c(45, 55, 65, 60, 50, 70, 85, 90, 35, 40)
mean(marks)
median(marks)
sd(marks)
skewness(marks)
kurtosis(marks)
```

Interpretation: Useful for understanding student performance spread.

6.2 Example 2: Daily Temperatures (in Celsius)

```
temps <- c(30, 32, 35, 33, 31, 29, 30, 36, 34, 32)
summary(temps)
sd(temps)
skewness(temps)
kurtosis(temps)
```

Interpretation: Helpful in studying weather consistency.

6.3 Example 3: Customer Ratings on Product (1 to 5 scale)

```
ratings <- c(5, 4, 4, 5, 3, 2, 5, 4, 1, 5)
mean(ratings)
mode <- get_mode(ratings)
sd(ratings)
skewness(ratings)
kurtosis(ratings)
```

Interpretation: Analyze satisfaction skew and peaks.

7 Case Study: Employee Salary Analysis

Problem: Analyze salary data for understanding income distribution in a company.

```
salaries <- c(30000, 35000, 40000, 38000,
41000, 70000, 75000, 80000, 100000, 120000)
summary(salaries)
mean(salaries)
median(salaries)
sd(salaries)
skewness(salaries)
kurtosis(salaries)
```

Insights:

- High standard deviation and positive skew due to few high salaries.
- Median less than Mean, indicating right skew.
- Kurtosis greater than 3 indicates sharper peak—income inequality exists.

8 Practice Questions

1. Use the 'cars' dataset to compute mean, median, and standard deviation of speed and distance.
2. Use the 'mtcars' dataset to analyze variability in MPG and horsepower.
3. Use the 'iris' dataset to find skewness and kurtosis of Sepal.Length.
4. Compare mean and median in 'cars' dataset to assess skewness.
5. Write a script that displays all descriptive statistics for any column of an inbuilt dataset.

9 Practice Questions-Solutions

1. Use the 'cars' dataset to compute mean, median, and standard deviation of speed and distance.

```
mean(cars$speed)
median(cars$speed)
sd(cars$speed)
mean(cars$dist)
median(cars$dist)
sd(cars$dist)
```

2. Use the 'mtcars' dataset to analyze variability in MPG and horsepower.

```
var(mtcars$mpg)
sd(mtcars$mpg)
var(mtcars$hp)
sd(mtcars$hp)
```

3. Use the 'iris' dataset to find skewness and kurtosis of Sepal.Length.

```
library(moments)
skewness(iris$Sepal.Length)
kurtosis(iris$Sepal.Length)
```

4. Compare mean and median in 'cars' dataset to assess skewness.

```
mean_speed <- mean(cars$speed)
median_speed <- median(cars$speed)
if (mean_speed > median_speed) {
  print("Right skewed")
} else if (mean_speed < median_speed) {
  print("Left skewed")
} else {
  print("Symmetric")
}
```

5. Write a script that displays all descriptive statistics for any column of an inbuilt dataset.

```
get_stats <- function(x) {
  cat("Mean:", mean(x), "\n")
  cat("Median:", median(x), "\n")
  cat("Variance:", var(x), "\n")
  cat("Standard Deviation:", sd(x), "\n")
  cat("Skewness:", skewness(x), "\n")
  cat("Kurtosis:", kurtosis(x), "\n")
}
get_stats(mtcars$mpg)
```