Practical Work (TP): Bivariate Statistics using R

Part I: Recall on R Data Structures

What is R?

R is a programming language and environment designed for **statistical computing and graphics**. It is widely used in data analysis, machine learning, and scientific research. R provides a large number of built-in statistical functions and excellent graphical tools.

Main Features

- Open-source and free.
- Rich collection of statistical methods (regression, classification, clustering, etc.).
- High-quality graphics for data visualization.
- Thousands of additional packages available on CRAN.
- Integrates with other languages such as C, C++, and Python.

Installing R

Step 1: Download R from the CRAN website: https://cran.r-project.org/
Choose your operating system:

- Windows: Download the `.exe` installer.
- macOS: Download the `.pkg` file.
- Linux: Install with `sudo apt-get install r-base`.

Step 2: Install RStudio (Optional but Recommended)
Download from: https://posit.co/download/rstudio/
RStudio provides an IDE with a script editor, console, plots, and package manager.

Running R

After installation, you can start R in two ways:

- Open the **R Console** directly.
- Use **RStudio** for a more user-friendly environment.

1. Variables

- Store values with `<-` or `=`:

```
x <- 5       # numeric
y = "hello"  # character
z <- TRUE    # logical
```

- Check type: `class(x)`, `typeof(y)`

2. Vectors

- 1D sequence of the same type:

```
v <- c(1,2,3,4)
seq(0, 10, by=2)   # 0,2,4,6,8,10
rep(5, times=3)    # 5,5,5
```

- Indexing:

```
v[1]     # first element
v[2:4]   # elements 2 to 4
v[-1]    # all except first
```

- Operations: element-wise (`v+1`, `v*2`)

3. Matrices

- 2D arrays (rows, columns):

```
m <- matrix(1:9, nrow=3, ncol=3, byrow=TRUE)
m[1,2]     # row 1, col 2
m[ ,1]     # first column
m[2, ]     # second row
```

- Operations: `rowSums(m)`, `colMeans(m)`, `t(m)`, `m %*% m`

4. Factors

- Represent categorical variables:

```
gender <- factor(c("M","F","M","F"))
levels(gender)   # "F" "M"
```

5. Data Frames

- Table-like structure (columns can have different types):

```
df <- data.frame(Name=c("A","B"), Age=c(20,22), Score=c(15.5,18.0))
df$Age     # column
df[1, ]    # row
df[ ,2]    # column by index
```

Part II: Guided Recall for Solving Exercises

This section recalls the most useful commands for solving the TP. They are hints, not full solutions.

1. Creating a Data Frame

```
df <- data.frame(Speed=c(...), Consumption=c(...))
```

2. Inspecting Data

```
head(df)
str(df)
summary(df)
```

3. Accessing Elements

```
df$Speed       # a column
df[1, ]        # first row
df[ ,2]        # second column
```

4. Basic Statistics

```
mean(df$Speed)
mean(df$Consumption)
var(df$Speed)
cov(df$Speed, df$Consumption)
cor(df$Speed, df$Consumption)
```

5. Plotting

```
plot(df$Speed, df$Consumption, pch=19, col="blue")
```

6. Regression

```
model <- lm(Consumption ~ Speed, data=df)
summary(model)
abline(model, col="red")
```

Tip: With these commands you can compute means, variance, covariance, correlation, and regression lines.

Part III: Exercises

Example 1: Plant Production

We study the relationship between the amount of fertilizer applied (x_i, in kg/ha) and the wheat yield (y_i, in q/ha):

Fertilizer x_i	0	50	100	150	200	250
Yield y_i	15	20	32	40	43	44

1. Enter the dataset into R and display it as a data frame.
2. Draw the scatter plot of (x_i, y_i). What kind of relationship do you observe?
3. Compute the marginal means $\bar{x}$ and $\bar{y}$.
4. Compute $V(x)$, $V(y)$, $\sigma(x)$, $\sigma(y)$ and $\text{cov}(x, y)$.
5. Find the regression line of Y on X and interpret it.
6. Compute the correlation coefficient r and comment on its value.

Example 2: Agricultural Sciences

We study the relationship between irrigation level (x_i, in mm/week) and tomato yield (y_i, in kg/m^2):

Irrigation x_i	10	20	30	40	50	60
Tomato yield y_i	2.5	3.0	4.2	5.0	5.4	5.6

1. Enter the dataset into R and display it as a data frame.
2. Draw the scatter plot of (x_i, y_i). What does it suggest about the relation?
3. Compute the marginal means $\bar{x}$ and $\bar{y}$.
4. Compute $V(x)$, $V(y)$, $\sigma(x)$, $\sigma(y)$ and $\text{cov}(x, y)$.
5. Fit the regression line of Y on X and write its equation.
6. Compute the correlation coefficient r and discuss whether irrigation strongly influences yield.