Biostatistics I: Introduction to R

Eleni-Rosalina Andrinopoulou

Department of Biostatistics, Erasmus Medical Center

✉ e.andrinopoulou@erasmusmc.nl

🐦 @erandrinopoulou

Erasmus MC
University Medical Center Rotterdam

Summary - Introduction

- ► **R** is a great tool to explore and investigate the data
- ► Several statistical methods can be performed with **R**
- ► It is important to understand the methods before applying them in **R**

How to use

R uses packages that perform specific tasks

- ► Install package only once
- ► Load package every time you open **R**

Summary - Basics

Basic functions

- ► `getwd()`, `setwd()`,
- ► `is.na()`,
  `is.finite()`,
  `is.null()`

Import/Export

- ► `read.csv()`, `write.csv()`
- ► `read.xlsx()`, `write.xlsx()`
- ► `read.table()`, `write.table()`

Save/Load

- ► `save()`, `saveRDS()`
- ► `load()`, `readRDS()`

Summary - Common Objects

Data types
- numeric
- character
- integer
- logical
- factors
- str(), mode()

Data structures
- c()
- matrix()
- array()
- data.frame()
- list()

Other
- ls(), objects()

Summary - Indexing and Subsetting

Vectors
- ► []
- ► [""] - for categorical variables

Matrices
- ► [,]
- ► [[]], []

Arrays
- ► [, ,]

Data frames
- ► [,]
- ► [[]], []
- ► $

Lists
- ► []
- ► [[]]
- ► $

Summary - Data Transformation/Exploration/Visualization

Transformation
- ▶ `round()`
- ▶ `factor()`
- ▶ `order()`
- ▶ `reshape()`

Exploration
- ▶ `mean()`, `sd()`
- ▶ `median()`, `IQR()`
- ▶ `table()`

Visualization
- ▶ `plot()`, `legend()`
- ▶ `hist()`
- ▶ `barchart()`
- ▶ `boxplot()`
- ▶ `xyplot()`, `ggplot()`
- ▶ `par()`

Summary - Correlation

► Pearson correlation: linear association

A relationship is linear when a change in one variable is associated with a proportional change in the other variable
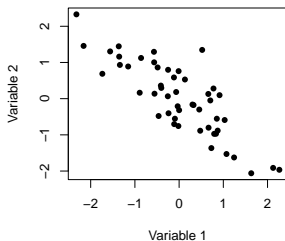
► Spearman correlation: monotonic relationship

A relationship is monotonic if the variables tend to change together, but not always at a constant rate
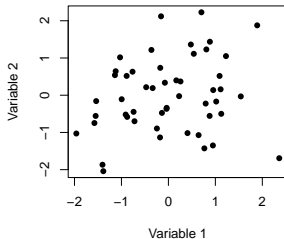

Positive correlation
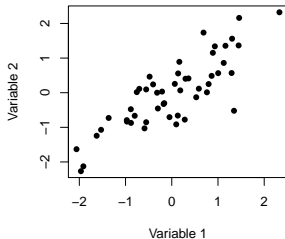
Negative correlation

No correlation

Summary - Hypothesis Testing

Types of tests

- ► Parametric / non-parametric
- ► one / two / M samples
- ► one-sided / two-sided

General procedure

- ► Choose a null hypothesis H_0 and an alternative hypothesis H_1
- ► Collect and visualize the data
- ► Choose and calculate the test statistic, which is a numerical summary of the data
- ► Determine the sampling distribution under the condition that the null-hypothesis holds
- ► Choose the type I error (significant level) α, usually α=0.05
- ► Determine the corresponding critical value(s)
- ► Compare the test statistic with critical value(s) and reject or not