



# Statistics with R: the basics

Master of Science in Signal Theory and Communications

TRACK: Signal Processing and Machine Learning for Big Data

**Instr.:** **J.L. Blanco-Murillo**

jl.blanco@upm.es, Room C-329

Departamento de Señales, Sistemas y Radiocomunicaciones

E.T.S. Ingenieros de Telecomunicación

Universidad Politécnica de Madrid

# Goal

- Introduce elementary functions / tools and learn about the language.
- Today we won't deepen into statistics.
- For now, focus is set on descriptive statistics, not on big data. Hence, we
  - avoid large-data optimization
  - limit resources required
  - grasp basic notions on the language and problems at hand

# Introduction (I)

- Descriptive Statistics: its aim is to present a data set in a way that is as informative as possible
- Different data nature
  - Qualitative. Non numerical characteristic
  - Discrete numbers. The observation result is an integer number
  - Continuous numbers. The observation result can be any value within some range

# Introduction (II)

- The techniques used in Descriptive Statistics can be
  - Numerical (centralization and dispersion measurements)
  - Graphical (bar plots, histograms…)

# Introduction (III)

- **E**xploratory **D**ata **A**nalysis
  - Uses descriptive statistics techniques to **clean** and analyze the data included in a **sample** that is studied to have an initial information about the whole **population**
  - Population: Whole set of elements from which we want to analyze something (get some knowledge or insights)
  - Sample: part of the population taken to make the analysis (usually, it is impossible to work on the whole population)
  - Example: Analyze the quality of the mobile phones made in a factory.

# Descriptive study of one variable

- The information we want to analyze, and present, depends on the **questions asked**, AND need to be **answered from the available data.**

- Usually, it is relevant to learn if…
  - there are concentrations in the data
  - there is a lot of dispersion
  - there is symmetry
  - there are “jumps” or missing places in the data range
  - there are outliers (atypical points)
  - there is a relation between the values of two data sets (more than one variable)