UP SKILLS

Co-funded by the
Erasmus+ Programme
of the European Union

A glimpse into language data science

Maja Miličević Petrović,
Adriano Ferraresi, Paul Marty

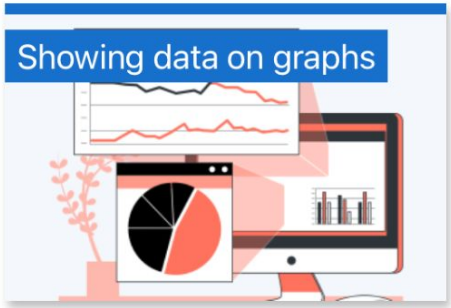
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

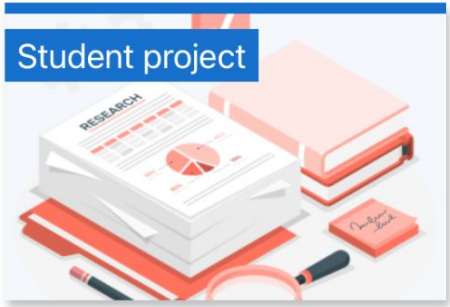
L-Università
ta' Malta

Guidelines and Best Practices for Research-Based Teaching

Utrecht, 4 November 2022

Outline





👩‍🏫 **Key concepts**

👨‍🎓 **Activities**

💾 **Data and code**

Formats

- **Theoretical and methodological contents**
  ⇒ Moodle books

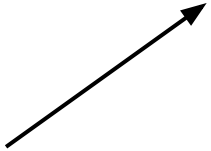


Table of contents

Some basic measurement-related distinctions include the quantitative/qualitative divide, also known as numerical/categorical, as well as the contiunous/discrete divide.

Quantitative or **numerical** variables are those whose values can be expressed numerically, such as age or word frequency. **Qualitative** or **categorical** variables are those whose values cannot be expressed numerically, but can instead be classified into categories, such as native language or part of speech.

Within quantitative variables, a further distinction can be made between **discrete** variables, whose values must be whole numbers (for example, the number of occurrences of a word in the corpus), and **continuous** variables, which that can take any value, be it a whole or a decimal number (for example, the time it takes to read a word).

The most widely used classification of measurement in social sciences is that in **four scales** defined by the psychologist Stanley Stevens, who distinguished between **nominal**, **ordinal**, **interval** and **ratio** scale. Many textbooks and tutorials rely on this division when explaining statistical procedures. Distinguishing between these scales is important because, mathematically speaking, some provide less and some more information about the phenomena they measure, leading to different choices of analyses.

The **nominal scale** comprises the values of categorical variables such as gender, part of speech, text type, or native language. These values can be expressed textually (noun, verb, adjective) or using numbers (noun -> 1, verb -> 2, adjective -> 3). In either case, the values do not have any real mathematical meaning and are used only for the purpose of classification. There is no relationship of numerical order or continuity between different values.

- **Exercises**
  ⇒ R scripts / R markdown

1 The case study
2 Importing data from files
3 Data preparation
4 Data treatment
5 Data analyses

```
# Read tabular data into R
## Read all text files
CaseStudy<-read.table("CaseStudy.csv", header = TRUE, sep=",")
## Read csv. files specifically
CaseStudy<-read.csv("CaseStudy.csv", header = TRUE, sep = ",")
```

Note that, for the above code to work properly, the file which the data is to be read from, namely "CaseStudy.csv", need to be in the current working directory, otherwise R will not know where to find this file on your computer in the absence of a complete file path.[2] If you encounter any issue, you may simply pass the function `file.choose()` to `read.table()` to open your file browser and search for the relevant file on your computer.

```
CaseStudy<-read.table(file.choose(), header = TRUE, sep=",")
```

Whichever method you use, the data should now be available in your current R working environment, stored in a data frame named `CaseStudy`. Before going any further, it is essential to inspect the structure and contents of this data frame to verify that the data file we imported into R was parsed as expected. For these purposes, we can use the following commands, which we have introduced in the previous unit:

(Re)use

The materials are...

- Downloadable
- Modifiable
- Modular

Different levels of (re)use possible:

- Full course
- Course units
- Books
  - Book chapters
  - Individual paragraphs
- Exercises