LAB: Univariate analysis

2026-01-30

M1 MIDS/MFA/LOGOS
Université Paris Cité
Année 2025
Course Homepage
Moodle

Univariate numerical samples

```
to_be_loaded <- c("tidyverse",
                  "magrittr",
                  "skimr",
                  "lobstr"
)

for (pck in to_be_loaded) {
  if (!require(pck, character.only = T)) {
    pak::pkg_install(pck) # ,,    repos="http://cran.rstudio.com/")
    stopifnot(require(pck, character.only = T))
  }
}
```

Objectives

In Exploratory analysis of tabular data, univariate analysis is the first step. It consists in exploring, summarizing, visualizing columns of a dataset.

In common circumstances, table wrangling is a prerequisite.

Then, univariate techniques depend on the kind of columns we are facing.

For *numerical* samples/columns, to name a few:

- Boxplots
- Histograms
- Density plots
- CDF
- Quantile functions
- Miscellanea

For categorical samples/columns, we have:

- Bar plots

- Column plots

Dataset

> Since 1948, the US Census Bureau carries out a monthly Current Population Survey, collecting data concerning residents aged above 15 from 150000 households. This survey is one of the most important sources of information concerning the american workforce. Data reported in file `Recensement.txt` originate from the 2012 census.

In this lab, we investigate the numerical colums of the dataset.

After downloading, dataset `Recensement` can be found in file `Recensement.csv`.

Choose a loading function for the format. `Rstudio` IDE provides a valuable helper.

Load the data into the session environment and call it `df`.

Table wrangling

> **Question**
>
> Which columns should be considered as categorical/factor?

Coerce the relevant columns as factors.

Search for missing data (optional)

> **Question**
>
> Check whether some columns contain missing data (use `is.na`).

> Useful functions:
>
> - `dplyr::summarise_all`
> - `tidyr::pivot_longer`
> - `dplyr::arrange`

Analysis of column `AGE`

Numerical summary

Use `skimr::skim()`

> **Question**
>
> Compare `mean` and `median`, `sd` and `IQR`.
> Are mean and median systematically related?

> **Question**
>
> Are standard deviation and IQR systematically related ?

Boxplots

Question

Draw a boxplot of the Age distribution

Question

How would you get rid of the useless ticks on the x-axis?

Histograms

Question

Plot a *histogram* of the empirical distribution of the `AGE` column

Question

Try different values for the `bins` parameter of `geom_histogram()`

Density estimates

Question

Plot a *density* estimate of the `AGE` column (use `stat_density`.

Question

Play with parameters `bw`, `kernel` and `adjust`.

Question

Overlay the two plots (histogram and density).

ECDF

Question

Plot the Empirical CDF of the AGE distribution

Question

Can you read the quartiles from the ECDF pplot?

Quantile function

Question

Plot the quantile function of the AGE distribution.

Repeat the analysis for `SAL_HOR`

Question

How could you comply with the DRY principle ?

Useful links

- veridical data science
- quarto
- rmarkdown
- dplyr
- ggplot2
- *R Graphic Cookbook*. Winston Chang. O' Reilly.
- A blog on ggplot object
- `skimr`