

TP n°1

stoehr@ceremade.dauphine.fr

Objectives This practical session is intended to teach you how to run and save commands in a script, import a data file, examine its contents and produce simple graphical representations. The final part illustrates concepts related to statistical testing, first with historical data and then through simulation.

Organising your work The distributed folder is organised as an R project. This is the recommended structure for your work. All commands will be run relative to the root of the project. Relative paths make the project easy to move from one computer to another.

1. After unzipping the folder, open the file `Statistical_Models.Rproj` with RStudio. Do not move the file `mais.txt` and do not use the function `setwd`.
2. In the R folder, create a script named `TP1.R`. Important commands should be saved in a script rather than run only in the console, so that the analysis can be reproduced. A line, or a selection of lines, can be run with `Ctrl + Enter` under Windows and Linux, or `Cmd + Enter` under macOS.
3. If it is not already available, install the package `ggplot2`.

```
install.packages("ggplot2")
```

4. Load the package `ggplot2`.

```
library(ggplot2)
```

Exercise 1.

1. Run the following commands from the script. What do they return?

```
getwd()  
list.files()  
list.files("data")
```

2. Create the following vector. The arrow `<-` is the assignment operator. Use the functions `length`, `mean`, `sd`, `median`, `min` and `max` to describe `x`.

```
x <- c(12, 15, 9, 18, 11, 14)
```

3. Compare the results of the following commands and explain the role of the square brackets.

```
x[1]  
x[2:4]  
x[x >= 14]
```

4. Open the help page for the function `mean` with the command `?mean`. Identify the argument used to ignore possible missing values. Save the script.

Exercise 2. The file `mais.txt` describes 288 experimental maize plots. For each plot, the data contain an identifier, a genetic marker, a variety, the plot yield and several characteristics of the grains and plants.

1. Import the data with the following command. Explain the role of the arguments `header` and `sep`.

```
mais <- read.table("data/mais.txt", header = TRUE, sep = "\t")
```

Then use `class`, `dim`, `names`, `str` and `summary` to examine the resulting object. Check the number of observations, the number of variables and whether any values are missing.

2. The variables `Marqueur` and `Variete` are qualitative. Convert them into factors using the function `factor`, and then use `str` again. Calculate the number of observations for each variety.
3. Display the columns `id`, `Variete` and `Rendement` for the first five plots. Then select the plots belonging to the `Tropical` variety. Calculate the mean, standard deviation and quartiles of the yield over all plots.

4. (a) Using base R, produce:

1. a histogram of the yield;
2. a bar chart of the numbers of plots for the different varieties;
3. boxplots of yield according to variety;
4. a scatter plot of the number of leaves against flowering time.

Give each graph an informative title and informative axis labels.

- (b) Briefly describe the main information provided by each representation and calculate the empirical correlation between `Floraison` and `Feuilles`.

5. (*Bonus*) Reproduce the four previous graphs with the package `ggplot2`.

Exercise 3. The dataset `morley`, supplied with R, contains 100 measurements made by Albert A. Michelson in 1879 to estimate the speed of light. The measurements are divided into five experiments of twenty runs. The variable `Speed` gives the measured speed after subtracting 299 000 km/s. The current reference value used in this exercise is $\mu_0 = 299\,792.458$ km/s.

1. Load the data with the following command.

```
data(morley)
```

Examine them using the tools introduced in Exercise 2. Convert `Expt` into a factor and create the variable `Vitesse`, expressed in km/s. Calculate the mean and standard deviation of the measurements for each experiment.

Student's t-test. We provisionally model the measurements as independent random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown. We wish to test

$$\mathcal{H}_0 : \mu = 299\,792.458 \quad \text{against} \quad \mathcal{H}_1 : \mu \neq 299\,792.458.$$

Under the null hypothesis, when the observations are independent and normally distributed, the test statistic

$$T_n = \frac{\sqrt{n}(\bar{Y}_n - \mu_0)}{S_n}, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2,$$

has a Student's t -distribution with $n - 1$ degrees of freedom. For a level $\alpha \in]0, 1[$, the rejection region of the two-sided test is

$$\mathcal{R}_n = \left\{ |T_n| > q_{1-\alpha/2}^{t(n-1)} \right\},$$

where $q_{1-\alpha/2}^{t(n-1)}$ denotes the quantile of order $1 - \alpha/2$ of Student's t -distribution with $n - 1$ degrees of freedom.

2. Perform the two-sided Student's t -test with `t.test` at the 5% level. In the output, identify the test statistic, the number of degrees of freedom, the p -value, the confidence interval and the estimate of μ . What decision is made? Explain why this conclusion should be interpreted with caution.

Illustration using simulated data. We now study the behaviour of the test in the controlled setting of synthetic data.

3. Take $n = 20$ and $\sigma = 80$ km/s, and use the random seed 2026 to simulate a sample under $\mathcal{H}_0$. Calculate the statistic and its two-sided p -value manually, and then check the results with `t.test`.
4. (a) Consider the sample sizes $n \in \{10, 20, 50, 100\}$ and the grid

$$\mathcal{G} = \{\mu_0 - 80, \mu_0 - 40, \mu_0 - 20, \mu_0, \mu_0 + 20, \mu_0 + 40, \mu_0 + 80\}.$$

For each n and each $\mu \in \mathcal{G}$, estimate the rejection probability of the two-sided Student's t -test at level $\alpha = 5\%$ by simulating $B = 1\,000$ samples of size n from the distribution $\mathcal{N}(\mu, \sigma^2)$, with $\sigma = 80$ km/s.

- (b) For $\mu = \mu_0$, compare this probability with the nominal level of the test. For $\mu \neq \mu_0$, the resulting probabilities estimate the power function

$$\mu \mapsto \pi_n(\mu) = \mathbb{P}_{\mu, \sigma}(\text{reject } \mathcal{H}_0), \quad \mu \neq \mu_0,$$

where $\sigma = 80$ km/s. Plot the results as a function of $\mu - \mu_0$, drawing one curve for each value of n . Comment on the influence of the sample size and of the difference between μ and μ_0 .

To go further.

5. (*Bonus*) Reproduce the following graph as accurately as possible in R. The function `stripchart` can be used to add points with a slight horizontal offset, while `adjustcolor` can be used to modify their transparency. Then reproduce the same graph with `ggplot2`. You may use `geom_boxplot`, `geom_point`, `position_jitter` and `geom_hline`.

Michelson's measurements in 1879