Descriptive statistics with R

At the end of this class, you should be able to :

- use R to carry out descriptive studies
- mobilize appropriate tools according to the nature of the data at hand
- interpret the results

1 Data and objectives

In this work, we will consider data from a randomized trial designed to study the effect of an anti-inflammatory drug (indomethacin) on the incidence of pancreatitis after a medical procedure (ERCP) aiming at diagnosing dysfunctions of some digestive organs [1]. The variables collected on each patient are as follows :

- `id` : subject id,
- `site` : study site (center),
  `1` = University of Michigan, `2` = Indiana University, `3` = University of Kentucky, `4` = Case Western,
- `age` : age in years,
- `risk` : risk score for post-ERCP pancreatitis,
- `gender` : male or female,
- `sod` : presence of sphincter of oddi dysfunction,
- `pep` : previous post-ERCP pancreatitis (PEP),
- `recpanc` : recurrent pancreatitis,
- `outcome` : outcome of post-ERCP pancreatitis,
- `status` : outpatient status,
- `type` : sphincter of Oddi dysfunction type/level - higher numbers are more severe,
- `rx` : treatment arm,
- `bleed` : a gastrointestinal bleed occurred.

Variables `sod`, `pep`, `recpanc` are known risk factors for pancreatitis and `bleed` is an adverse event of the intervention. The objective of this exercise is to make a detailed descriptive study of the sample.

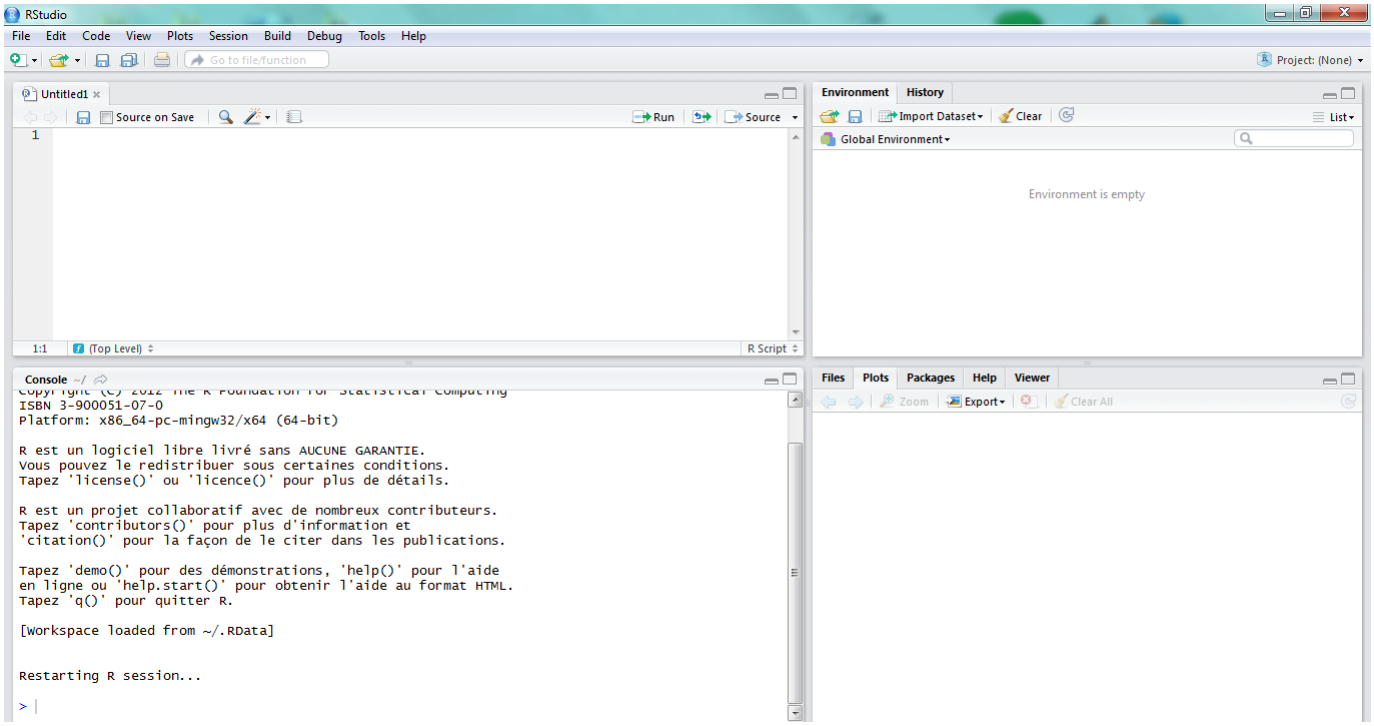
1. The original data is available in the `medicaldata` package under the name `indo_rct`. It has been deliberately simplified and extracted into a csv file for the purposes of this exercise. You can find out more about the original data structure by consulting the help.

2 Quick introduction to R and RStudio

2.1 Overview

R is a statistical software that can be downloaded for free at the following address : `http://www.r-project.org/` and installed under windows, unix or MacOS. It consists of a basic kernel and multiple packages developed and made available to all by users. **RStudio**, which can also be downloaded free of charge at the following address : `http://rstudio.org/`, provides a very user-friendly graphic interface. We will work exclusively under **RStudio**.
RStudio interface is divided into several windows :

1. **The editor** allows to write code and save it,
2. **The console** is used to run the code. Results other than graphs will also appear in this window.
3. The third window contains the workspace and the history of the commands,
4. The last window contains the tabs Files-Plots-Packages-View.



2.2 The editor

The editor available in **RStudio** allows you to write the code and save it in a file. Saving codes in files is particularly useful for making corrections, keeping track of your work and restarting programs after interruptions.

- If the editor does not appear when you open **RStudio**, you can create it from the menu `File - New File - R Script`.
- To save the contents of the editor, follow this path in the menu : `File - Save as`.

⚠**It is strongly advised to write instructions in the editor before executing them in the console and to save the contents of the editor regularly during the session**.

3 Data processing

3.1 Importing data into R

3.1.1 Data

1. The data are available in the form of a csv file `Data_indo.csv`
2. Once the file is saved in a directory of your choice, open it with a simple text editor (WordPad, Notepad, TextEdit ...) You'll notice that the file is organized in lines, where each line corresponds to an individual and always lists the values measured for each variable in the same order. Here, we can observe that the passage from one variable to the next is marked by a comma, and that the first line of the file is not an observation but the list of the variables names. Close the file.

3.1.2 Working directory

It is convenient to set a working directory where the data is located and where the results and program files will be saved. You can use the menu `Session - Set working directory - choose directory`.

3 – Specify your working directory into R .

3.2 Importing data

R is not able to work on the initial csv file. It is necessary to import the data from the csv file into a data table in R specific format before starting the analysis.

4 – Import data into variable `tab` with the following instructions

```
library(tidyverse)
tab <- read_csv("Data_indo.csv",col_names = T)
```

To do this, place yourself in the editor and write the instruction there. To execute it, select the command, then simultaneously press the `Ctrl` and `Enter` keys on the keyboard. The function `read_csv` is part of a library that is not installed by default in R . You must therefore load this library beforehand using the instruction `library(tidyverse)`.
⚠R **is case-sensitive!**

5 – What do you think is the role of the option `col_names=T` in the data import instruction `read_csv`? You can also find this information in the help of R by running the command

```
help(read_csv)
```

or

```
?read_csv
```

3.3 Checking data importation

Let's now check that data importation has been carried out correctly.

6 – Run the commands `tab`, `head(tab)` and `str(tab)` successively. What do you get in each of the three cases?

7 – How many observations (rows) and how many variables (columns) are there in `tab`?

8 – What is the nature (quantitative, qualitative) of each variable? In the case of qualitative variables, specify the number of categories.

9 – Should variable `bleed` actually be numeric?

10 – Run the command :

```
tab$bleed <-  as.factor(tab$bleed)
```

What is the difference with the original data? To answer this question, you can use the instruction `str()`.
Note that the `$` sign in `tab$bleed` gives access to the variable `bleed` of the `tab` object. You can try running the commands `bleed` and `tab$bleed` and observe the differences.

11 – Check that the formats of the other variables coincide with their true nature and make the appropriate corrections if necessary.

3.4 Univariate descriptive statistics

12 – Recall the nature of variable `age`. In your opinion, what statistical tools may be relevant to describe this variable?

13 – To obtain a summary of the information contained in variable `age`, execute the following instructions :

(a) `summary(tab$age)`

(b) `sd(tab$age)`

(c) `min(tab$age)`

(d) `max(tab$age)`

(e) `mean(tab$age)`

14 – To which informations do these instructions give access to? Comment on the results.

15 – We then give a graphic representation with :

```
ggplot(tab, aes(x=age)) + coord_flip() + geom_boxplot()
```

(a) Recall the name of such a graphic.

(b) What do the different components of this graph represent?

(c) Comment on the appearance of the resulting graph.

16 – The following instruction leads to another interesting representation

```
ggplot(tab, aes(x=age)) + geom_histogram() + xlab("Age")
```

Recall the name of such graph and comment on its shape.

17 – Recall the nature of variable `type`. Which tool(s) do you feel would be appropriate for describing this variable?

18 – Execute the following instruction

```
summary(tab$type)
```

What is the result? Why is it different from the result obtained by applying the same function to variable `age`?

An equivalent result can be obtained with

```
table(tab$type)
```

19 – Some graphical descriptions of variable `type` can be obtained with the following instructions :

```
# Bar plot of the counts in each category
ggplot(tab, aes(x=type)) + geom_bar() + xlab("Dysfunction type")
# Bar plot of the proportions in each category
tab.type <- tab %>% count(type) %>%
mutate(perc = n / nrow(tab))
ggplot(tab.type, aes(x = type, y = perc)) + geom_bar(stat = "identity")
# Pie chart
ggplot(tab.type, aes(x = "", y = perc,fill=type)) +
geom_bar(stat = "identity",width=1) + coord_polar("y", start=0)
```

Comment on the graphs.

20 – Provide a descriptive study of variable `outcome` that measures the event of post-ERCP pancreatitis in patients involved in the clinical trial. Comment on the results.

3.5 Bivariate descriptive statistics

It is common in such clinical study to generate descriptive statistics by treatment arm in order to compare the characteristics of patients assigned to each group. The treatment arm is given by variable `rx` (0 : placebo, 1 : treatment).

21 – Using the instructions introduced earlier, how many patients are there in each treatment group?

22 – We first want to compare the ages of the patients in the two groups.

(a) Recall the nature of variables `age` and `rx`

(b) Execute the instructions

```
by(tab$age,tab$rx,mean)
by(tab$age,tab$rx,summary)
```

What is the purpose of the `by` instruction? Comment on the results and complete this bivariate descriptive study with other indicators.

(c) Propose a graphical representation to compare the ages of the patients in the two groups.

23 – We now seek to compare the occurrence of pancreatitis between the two treatment arms.

(a) What is the nature of the variables of interest in this comparison.

(b) Run the following instructions :

```
table(tab$rx,tab$outcome)
proportions(table(tab$rx,tab$outcome),margin=1)
```

What does each of these instructions do? What do you conclude from the results?

(c) One can also generate the following graph :

```
tab.prop <- proportions(table(tab$rx,tab$outcome),margin=1)
data.prop <- data.frame(pct = as.vector(tab.prop), rx = rep(levels(tab$rx),
length(levels(tab$outcome))), outcome = rep(levels(tab$rx),
each=length(levels(tab$outcome))) )
ggplot(data.prop, aes(x=rx, y=pct, fill=outcome)) +
geom_bar(stat="identity", position=position_dodge())
```

What does this graph represent? Comment it on.

24 – The score given in variable `risk` measures a preoperative risk of pancreatitis based on patient characteristics. For simplicity, patients with a score less than or equal to 2 are considered low risk and patients with a score greater than 2 are considered high risk.

(a) What does the following instruction do?

```
tab <- tab %>% mutate(risk_bin = (risk>2))
```

(b) What do the three tables generated below tell us about?

```
proportions(table(tab$risk_bin,tab$outcome),margin=1)
tab_placebo <- tab %>% filter(rx=="0_placebo")
proportions(table(tab_placebo$risk_bin,tab_placebo$outcome),
margin=1)
tab_treat <- tab %>% filter(rx=="1_indomethacin")
proportions(table(tab_treat$risk_bin,tab_treat$outcome),
margin=1)
```

4 Personal work

Complete the descriptive study by examining the other variables at hand.