Manipulating datasets with R (lesson 2)

Emmanuelle Becker, University of Rennes 1

Contents

1 Manipulating datasets 1
  1.1 The read.table() function ........................................... 1
  1.2 Exploring the table .................................................. 2
  1.3 Columns selection .................................................... 2

2 Introduction to graphics 3
  2.1 General Comments ................................................... 3
  2.2 First graphs .......................................................... 3

1 Manipulating datasets

1.1 The read.table() function

This function allows to import a dataset contained in a text file or a spreadsheet. The columns
are the variables studied, and the lines are the individuals of the population (exactly one indi-
vidual per line). The supported format is very simple : columns are separated with a comma, a
space or a tabulation (or anything we want if we specify it with the \texttt{sep} argument). The resulting
object in R is called a \texttt{data.frame}.

\begin{verbatim}
# reading the table ...
> tab = read.table("data-individuals.txt", header=TRUE, sep="\t")

# looking at the table that has been read (to detect potential errors)
# with at least one of these three commands
> head(tab)
> str(tab)
> summary(tab)
\end{verbatim}

Very important #1 If the command does not work, in 99% of the cases it is because (i) the
filename is not correct ; or (ii) the working directory is not correct.

What is the working directory ? This is the directory where R is searching for your data
files. Everyone has its own way to organize its files and directories in its computer, so R cannot
guess in which directory are your files. You thus have to indicate it ! To indicate your working
directory, use the menu Session > set working directory > choose directory.

Very important #2 When you have loaded a dataset, always verify the import is correct. To do
that, you can use the command \texttt{head}, like : head(tab)
What is a data.frame exactly? A table of data, of type data.frame is a list of data vectors:

- all vectors are the same length
- some are numeric, others logics, other factorials etc.
- vectors can be named (names of variables studied)
- lines can be named (labels of observations)
- it can be manipulated as a matrix, with 2 indices \( \text{data}[i,j] \), but it is not a matrix.

1.2 Exploring the table

How many lines and columns contain this table (functions \( \text{dim()} \), \( \text{nrow()} \), \( \text{ncol()} \))?

What are the data contained in the table (functions \( \text{names()} \), \( \text{summary()} \))?

1.3 Columns selection

To access the content of an entire column, two solutions are possible:

1. The first is to use the name of the column (if there are headers of course). For example, to access the vector of weights:
   > \text{tab$Weight}

2. The second method is to give the number of the column. For example, the weight being the third column, we can access the vector of weights with
   > \text{tab[,3]}

Exercise 1: Calculate the average weight, and the standard deviation of weight.

> \text{mean(tab$Weight)}
> \text{sd(tab$Weight)}

Exercise 2: Calculate the average height, and the standard deviation of height.

Exercise 3: Calculate the average size of women (the command is given), then that of men.

> \text{mean(tab$Height[tab$Gender=='Mrs'])}
...

Exercise 4: How many women are taller than 1m70 with blue eyes?

Exercise 5: What is the average weight of men wearing glasses?

Exercise 6: How many men are wearing glasses? (command below)

> \text{sum(tab$Gender=='Mr' & tab$Glasses=='Yes')}

Exercise 7: What is the median weight of men studying Economy?

Exercise 8: How many men are studying Economy?
Exercise 9 (difficult): What is the weight of the thinnest man? Does he wear glasses?

Exercise 10: Calculate the body mass index (bmi), for all the individuals in the table (the formula is given below).

\[ bmi = \frac{\text{weight}}{\text{height}^2} \]

(the size is in m, the weight in kgs)

2 Introduction to graphics

2.1 General Comments

R is a very good environment to produce high quality graphics. When exploring a dataset, using a good graphical representation is a key point. We are going to present them very quickly (they will be more detailed in the next practical sessions).

2.2 First graphs

In RStudio, the figures are displayed in the window at the bottom right, and you can save them easily by clicking on Export. Many formats are offered.

Exercise 11: Choose one function between `hist(X)`, `barplot(table(X))`, `plot(X, Y)`, and `boxplot(Y ~ X)`, to draw the diagrams giving:

- the distribution of sizes;
- the distribution of major domains studied;
- a box-and-whisker type diagram with size according to gender;
- a scatter plot showing size by weight.

Exercise 12: Try the command:

```r
> barplot(table(tab$Major), col=c("blue","magenta","lime"),
        main="Major Choice",
        xlab="Domains", ylab="Number")
```

and explain what are the arguments `main`, `xlab`...