

1- Introduction to the R language

Alex Sanchez, Miriam Mota and Santiago Perez-Hoyos

Statistics and Bioinformatics Unit. Vall d'Hebron Institut de
Recerca

Section 1

Introduction to R

Outline

- A first contact with R & Rstudio.
 - How does one work with R
- A primer of data import
 - Reading data into R
- A primer of communication report
 - R Notebooks and RMarkdown

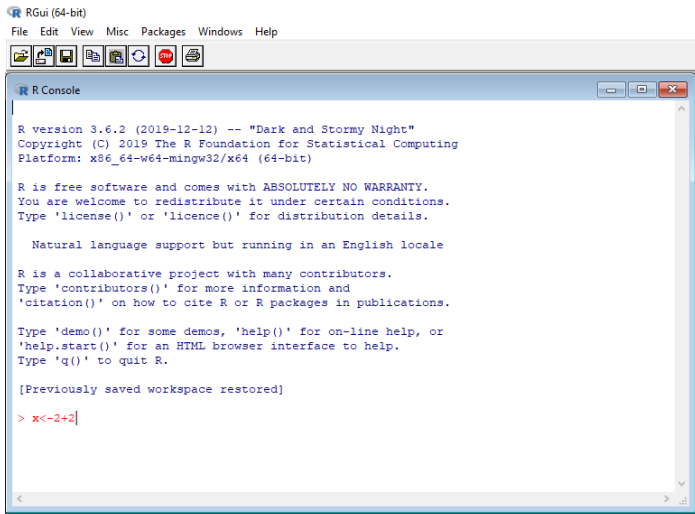
Key Terms

- **R** is a *language and environment* for statistical computing and graphics.
- **R Studio** Graphical User Interface for easier use of R
- **Objects** Everything you store in R (datasets, variables, lists, results and graphs outputs) that can be referenced and reused
- **Functions** Pre-built lines of code that execute actions after inputting some parameters.
- **Packages or library** Shareable bundle of code and documentation that contains pre-defined functions. R contains base packages and for some analysis you must install and call specific ones.
- **Scripts** Document file that hold your commands that can be run later.
- **Rmarkdown** Special type of Script that can mixed text and comments with R commands that can be compiled in a final pdf,

How is R used

- Traditionally R was used from an Operating System console (“Terminal”)
- This is an intimidating approach for many users
- A variety of options exist to decrease the learning curve.
 - Use a supportive development environment such as **Rstudio**
 - Use an interface to Statistical tools with menus, such as **Rcommander** or **Jamovi** allowing to concentrate on Statistics, not in commands.

A raw R console



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> x<-2+2|
```

An “enhanced” console: Rstudio

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code for loading libraries, reading data, and fitting a survival model.
- Console:** Shows the execution output of the R code, including the results of `summary(KM_Fit)` and `ggsurvplot`.
- Environment Pane:** Lists the objects in the global environment, including `dat2`, `diabetes`, `diabetes_factor`, `tbl_df`, and `KM_fit`.
- Plot Pane:** Displays a Kaplan-Meier survival plot comparing the survival probability over time for two groups: `Strata` (red line) and `All` (black line).

Script Editor Code:

```
10 editor_options:
11   chunk_output_type: inline
12
13
14
15
16 [r]
17 library(readxl)
18 library(dplyr)
19 library(ggplot2)
20 library(ggfortify)
21 library(survival)
22 library(survminer)
23 library(gsumsummary)
24
25
26
27 lectura de los datos
28 [r]
29 diabetes <- read_excel("datasets/diabetes.xls")
30 sapply(diabetes, class)
31 diabetes_factor <- diabetes %>%
32   mutate_if(sapply(diabetes, is.character), as.factor) %>%
33   select(-numparts)
34 sapply(diabetes_factor, class)
35 dat2 <- diabetes_factor
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```

Console Output:

```
## Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ 1, data = dat2)
##
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
## 0 149 0 1.000 0.00000 1.000 1.000
## 2 142 1 0.993 0.00702 0.979 1.000
## 4 133 3 0.972 0.01404 0.944 0.999
## 6 123 3 0.949 0.02080 0.913 0.987
## 8 110 6 0.901 0.02609 0.851 0.954
## 10 102 2 0.884 0.02837 0.830 0.941

## Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ 1, data = dat2)
##
## time n events median 0.95LCL 0.95UCL
## 0 149 0 1.000 0.00000 1.000 1.000
## 2 142 1 0.993 0.00702 0.979 1.000
## 4 133 3 0.972 0.01404 0.944 0.999
## 6 123 3 0.949 0.02080 0.913 0.987
## 8 110 6 0.901 0.02609 0.851 0.954
## 10 102 2 0.884 0.02837 0.830 0.941

## Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ 1, data = dat2)
##
## time n events median 0.95LCL 0.95UCL
## 0 149 0 1.000 0.00000 1.000 1.000
## 2 142 1 0.993 0.00702 0.979 1.000
## 4 133 3 0.972 0.01404 0.944 0.999
## 6 123 3 0.949 0.02080 0.913 0.987
## 8 110 6 0.901 0.02609 0.851 0.954
## 10 102 2 0.884 0.02837 0.830 0.941
```

Environment Pane:

Name	Type	Length	Size	Value
dat2	tbl_df	10	0 B	149 obs. of 10 variables
diabetes	tbl_df	11	15.5 KB	149 obs. of 11 variables
diabetes_factor	tbl_df	10	0 B	149 obs. of 10 variables
KM_fit	survfit	16	10.3 KB	List of 16

Plot Pane:

summary(KM_fit, times=c(0,2,4,6,8,10))

Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ 1, data = dat2)

##

time n.risk n.event survival std.err lower 95% CI upper 95% CI

0 149 0 1.000 0.00000 1.000 1.000

2 142 1 0.993 0.00702 0.979 1.000

4 133 3 0.972 0.01404 0.944 0.999

6 123 3 0.949 0.02080 0.913 0.987

8 110 6 0.901 0.02609 0.851 0.954

10 102 2 0.884 0.02837 0.830 0.941

ggsurvplot(KM_fit, data = dat2, risk.table = TRUE)

Strata: All

1.00

0.75

0.50

ai probability

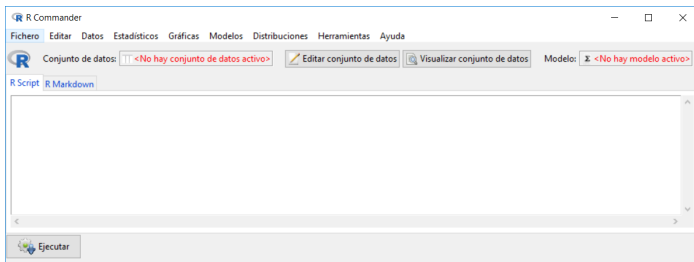
An “enhanced” console: Rstudio

The image shows the RStudio interface with several key components highlighted by colored boxes and arrows:

- Source: Script Edition and Run:** Points to the top-left pane where R code is written. It includes a toolbar with icons for running code (a green play button) and inserting chunks (a green plus icon).
- Environment objects:** Points to the top-right pane, which displays a table of objects currently loaded in the R environment.

Name	Type	Size	Value
dat2	tbl_df	10	0 8 149 obs. of 10 variables
diabetes	tbl_df	11	15.5 KB 149 obs. of 11 variables
diabetes_factor	tbl_df	10	0 8 149 obs. of 10 variables
RM_fit	survfit	16	10.3 KB List of 16
- R Console:** Points to the bottom-left pane, which shows the execution history and output of the R code. The code includes loading libraries (readr, dplyr, ggplot2, plotly, survival, rstanarm), reading data from an Excel file, and performing survival analysis using the `survfit` and `ggsurvplot` functions.
- Plots, Packages and Help:** Points to the bottom-right pane, which displays a Kaplan-Meier survival plot. The plot shows the probability of survival over time, with a red line representing the estimated survival function and a shaded area representing the confidence interval. The x-axis is labeled 'Time' and the y-axis is labeled 'Probability'.

Something that is not a console: Rcommander



Exercise

- Open R-Studio
- Identify Panes in R
- Calculate $2+2$ in the console

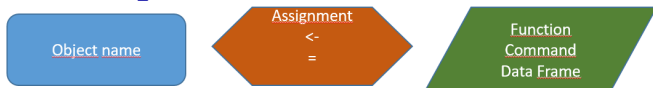
Open new Script

- Run $2+2$ Command line in Script

Section 2

Using R

Commands, Objects and Functions



Examples

- Vector

```
x1<-c(3,4,6,9,12)
```

```
x2<-c(3,4,6,9,20)
```

- Data Frames

```
dades<-data.frame(x1,x2)
```

- Results of execution of Functions

```
summary(x1,dat=dades)
```

Exercise

- Create a Script with the commands of the previous slide and see the results

```
x1<-c(3,4,6,9,12)  # Create vector
x1  # Show vector
x2<-c(3,4,6,9,20)  # Create vector
x2  # Show vector
dades<-data.frame(x1,x2)  # Create database
dades  # Show database
summary(x1,dat=dades)  # Summary measures from database
```

Exercise

```
x1<-c(3,4,6,9,12) # Create vector  
x1 # Show vector
```

```
## [1] 3 4 6 9 12
```

```
x2<-c(3,4,6,9,20) # Create vector  
x2 # Show vector
```

```
## [1] 3 4 6 9 20
```

```
dades<-data.frame(x1,x2) # Create database  
dades # Show database
```

```
##   x1 x2  
## 1  3  3  
## 2  4  4  
## 3  6  6  
## 4  9  9  
## 5 12 20
```

```
summary(x1,dat=dades) # Summary measures from database
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
##       3.0     4.0     6.0     6.8     9.0    12.0
```

Section 3

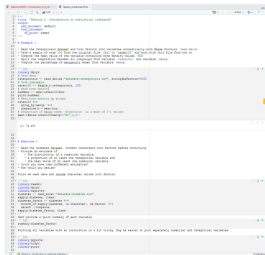
Dynamic output with Rmarkdown

Reproducible research with Rmarkdown

- R and Rstudio are strongly involved in promoting reproducibility and reproducible research.
- This is implemented in **Rmarkdown**
- A Rmarkdown combines
 - Natural language text, e.g. describing what we are doing in our own words.
 - R code with the instructions needed to do the data management or the analysis.
 - The output of the analysis

Reports in Rmarkdown

Rmarkdown Script



Creating Rmarkdown

- A Rmarkdown can be created in Rstudio with
 - File --> New File --> Rmarkdown
- The Rmarkdown contains example text and code so it is straightforward to adapt it to your analysis.
- To produce an html file with text, code and output:
 - Press the button “Knitr to Html”

Rmarkdown Script

```
StatisticsWithR-1-Introduction_to_R_cb... x  Untitled1 x
1 ---
2 title: "Demo Rmarkdown"
3 author: "UEB"
4 date: "27/5/2021"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF,
15 and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
16
17 When you click the Knit button a document will be generated that includes both content as
18 well as the output of any embedded R code chunks within the document. You can embed an R code
19 chunk like this:
20
21 ```{r cars}
22 summary(cars)
23 ```
24
25 ## Including Plots
26
27 You can also embed plots, for example:
28
29 ```{r pressure, echo=FALSE}
30 plot(pressure)
31 ```
32
33 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R
34 code that generated the plot.
```

YAML

R Chunk

Markdown

R Chunk

Markdown

R Chunk

Markdown

Exercise

- Create a new Rmarkdown document
 - Include a title and your name
 - Compile document with 'knitr to html'

R packages

- R can be used for many different types of data processing and analysis from distinct fields, besides statistics such as Ecology, Omics Sciences, Psychology etc.
- All these capabilities are not present from the beginning because most of them will never be used by most users.
- Instead, they can be added when needed by
 - (i) installing and
 - (ii) loading the appropriate packages.

Installing and loading packages

We want to analyze some data using cox proportional hazards model.

```
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)
```

```
Error in coxph(Surv(time, status) ~ sex, data = lung)  
: could not find function "coxph"
```

We need to install and load the package before we can use it.

```
install.packages("survival")  
library(survival)  
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)
```

The tidyverse

- The tidyverse is an opinionated collection of R packages designed for data science.
- All packages share an underlying design philosophy, grammar, and data structures.
- The complete tidyverse collection can be installed with:

```
install.packages("tidyverse")
```

- <https://www.tidyverse.org/>

Exercise

- Install the rio package from menu.
- Load the rio package.

Section 4

Getting data into R

Importing data with Rstudio

- The easiest way to get data into R is to click on the Import Datasets button.
- Alternatively R code can be written using functions from Base R, the tidyverse or rio package
 - Base R functions start with `read.:` `read.table`, `read.csv`
 - tidyverse functions start with `read_:` `read_delim`, `read_csv` or `read_excel`
 - rio function is `import`

Reading Excel or csv files

- Files can be read from any location, let it be a physical support or a web site.
- To read files from disk be sure to indicate their location.
- Alternatively the default working directory can be set to the folder where the file is located.
- Assume files `Diabetes.xls` and `Osteoporosis.csv` have been downloaded from url to a sub-folder named `datasets`
- Start setting the default directory to the folder where you have saved the `datasets` folder.
 - Session --> Set Working directory --> To source file location...
- Import the `diabetes.xls` and the `osteoporosis.csv` file

Reading Excel or csv files (continued)

The code generated for reading the files can be reused any time changing the file name if needed.

```
# Read Excel file  
library(readxl)  
diabetes <- read_excel("../datasets/diabetes.xls")
```

Reading text files

- Text files may require that more information is provided about delimiters, decimal sign, locale (language) or page encoding (UTFB for Mac or Linux vs ISO-8859-1 for Windows).
- All options can be selected from the rstudio importer

```
library(readr)
osteoporosis <- read_delim("../datasets/osteoporosis.csv",
  "\t", escape_double = FALSE, locale = locale(date_name = "full",
  decimal_mark = ",", encoding = "ISO-8859-1"))
```

Reading Excel or csv files with rio

```
require(rio)
import("../datasets/diabetes.xls")
import("../datasets/osteoporosis.csv", dec = ",")
```

Interlude: Summarizing data

- Once a dataset is available it is easy to “have a look at it”

```
head(diabetes)
str(diabetes)
dim(diabetes)
summary (diabetes)
```

Section 5

Resources and exercises

Introductory materials

The web is full of all types of materials about R

Below there are a couple of brief introductions:

- A short introduction to R
- Getting started with R

Exercise

- Select a dataset with which you wish to work along the course.
- Read it into R
 - How many variables are there in it
 - What are their types
- Try to summarize it briefly
- Create an Rmarkdown to encapsulate all your steps and share it with somebody.