Introduction to R, RStudio

MF9130E – Introductory Course in Statistics
25.04.2023

Chi Zhang
chi.zhang@medisin.uio.no
Oslo Center for Biostatistics and Epidemiology
Department of Biostatistics, UiO

About

About me:
PhD in Biostatistics, UiO. Currently researcher / R developer. R user since 2015

<u>About the guided R lab sessions:</u>

Aim: help you get started with R, so that you can use it for your own statistical analysis

Week 1 is scheduled in this way:

Afternoon: **theory**
Morning after: **theory recap, examples, practice**

In week 1, we will cover topics such as descriptive statistics, probability distributions, hypothesis testing, comparing continuous variables, categorical data analysis

Outline

8:30-9:30	Introduction to R and RStudio
Break	
9:45-10:45	Recap theory: descriptive statistics
	Guided practice (exercise 1, 2)
Break	
	Practice
11:15-11:30	Summary and wrap up

Lab notes for today:
(under *R Lab and Code* tab)

Getting started in RStudio

Introduction to R

EDA I

Link to *R Lab and Code*

https://ocbe-uio.github.io/teaching_mf9130e/lab/overview.html

Why we use R in this course?

R has a few advantages over other softwares (e.g. STATA)

- It implements not only the **classical statistical models**, but also **latest developments** in methodology and computation
- Free, no need for license
- Powerful and flexible visualization tools
- Lots of users in different fields; active community: you can get help easily online
- In the future, your statistician is more likely to know R than other softwares

RStudio interface

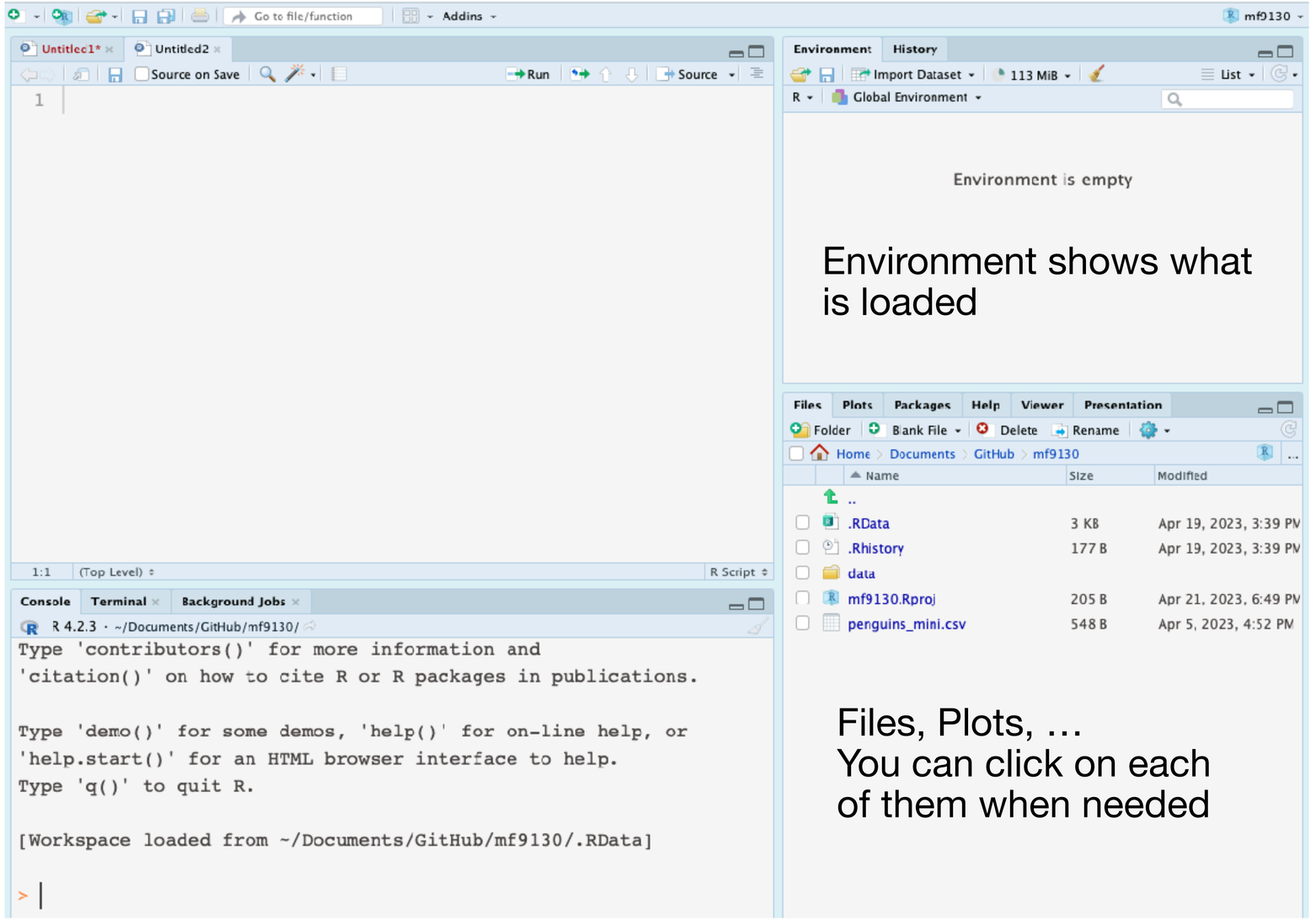
Script (source)

You can add commands and save them;

RUN the commands line by line with ctrl+enter (cmd+enter in macOS)

Console

This is where R code is executed (run), and returns results



In this session

How to run a command in R/Rstudio

Create a variable

Data types (numeric, character, logical)

Data structure (vector, matrix, data frame)

(*basic data manipulation*)

Import a dataset (R project, work directory)

90% of the work is **data cleaning and manipulation**!

(In which you gain a better understanding of your data)

Create a variable (in console)

Open RStudio

Locate **Console**

Type in `a <- 3` in Console (after >)

Click **enter**

Create a variable (use R script)

Create an empty R **script**, call it “first_script.r”

In the script, type in `b <- 5`

Execute this line with **ctrl+enter** (**cmd+enter**)

R as calculator

```
a <- 3
b <- 4
c <- 7

# calculate the average of a,b,c

(a+b+c)/3
```

Data types

Common data types:

Numeric - numbers, such as 1.2, -1

(‘double’, ‘integer’…)

Character - “hadley”, “female”

(‘string’)

Logical - true or false, 1/0

(‘binary’)

```
a <- 3.1
class(a)

student <- “hadley”
class(student)

true_or_false <- T
# or TRUE
class(true_or_false)
```

Data structure

Scalar

33

Vector

1	2	3	4

a
b
c
d
e

Matrix

1	2	3
4	5	6
7	8	9

Data frame

age	sex	Smoker
33	F	FALSE
44	M	TRUE
34	M	FALSE

The elements in vectors and matrices need to be of the same type: all numeric; all characters, etc

Each **column** of a data frame needs to be of the same type

Different columns can be of different types

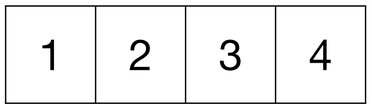
Your data most likely have mixed types of variables

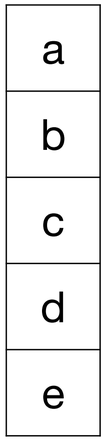
Vector

Create a vector

Use c() to combine elements (scalars)

Some shortcuts to create a sequence





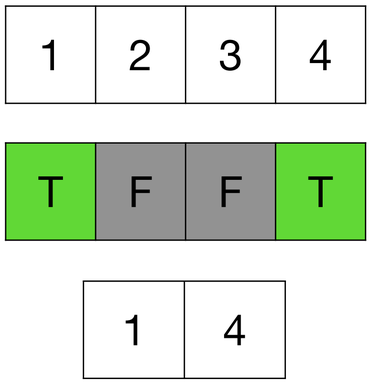

```
v1 <- c(1, 2, 3, 4)
v2 <- c("a", "b", "c", "d", "e")

# shortcuts

rep(0, 5) # repeat 0 for 5 times
v20 <- 1:20 # from 1 to 20

# combine with math operation

v20 * 2
```

Vector

Select elements of a vector

Select with indices (e.g. first, 3rd)

Select based on logical vector




```
# select the 3rd elements of v2
v1[3]


# select first 10 elements of v20
v20[1:10]

# select based on vector (T,F,F,T)
# prints 1 and 4
v1[c(T,F,F,T)]
```

(you will need it to filter two variables:
e.g. select heights for men and women (filter based on sex))

data.frame

Each row is a **subject**
(usually with unique IDs)

Each column is a **variable**
(feature, measurement, parameter)

age	sex	smoker
33	F	FALSE
44	M	TRUE
34	M	FALSE

Select elements from data frame
Index
$ operator
Variable name

```
# select the 3rd elements of v2
df <- data.frame(
  age = c(33, 44, 34),
  sex = c(“F”, “M”, “M”),
  smoker = c(F, T, F)
)
df

# select first subject
df[1, ]

# select variable age
df$age
df[, “age”]
df[[“age”]] # not df[“age”]
```

Working directory and R project

We are almost ready to practice the skills on a dataset!

Before we start the analysis, better keep organized!

Do you know where you are keeping the data, and where to tell R to look for it?

Working directory and R project

Load penguin_mini.csv data

(Let's download it together)

File system in your laptop

~/folder/sub_folder/…/file

R project can help make your work more organized!

```
getwd() # know where you are

# load a csv file (absolute path)
read.csv("~/Documents/your_user_name/
data_folder/another_folder/
penguin_mini.csv)


# load from R project (absolute)
read.csv("~/Documents/this_project/data/
penguin_mini.csv")

# load from R project (relative)
read.csv("data/penguin_mini.csv")
```

Working directory and R project

You are ready to carry out analysis on datasets!

We continue in the next session, on how to explore your data in R.

We will use two slightly larger datasets.