Exercises Day 1*

Ronald Geskus Tran Thai Hung

Last version: April 04, 2025

You already made exercises 1 and 2 in the slides. For exercises 3 to 7, you will do them in RStudio. Please make sure that you have R and RStudio installed on your laptop.

Table of contents

Exercise 1: selection within a vector

a. Create a vector named `vec` that consists of the numbers from 11 to 30
b. Select the 7th element of the vector
c. Select all elements except the 15th. Hint: use a minus sign
d. Select the 2nd and 5th element of the vector
e. Select only the odd valued elements of `vec`.

*Please do not circulate.

Remark: Exercise e. is not an easy exercise. We advise you to do this is in two steps. First create a vector ***`index`*** *that selects the odd numbers in* ***`vec`****, using the function* ***`seq`****. You may need to consult the help page of* ***`seq`*** *via* ***`help(seq)`****.*

Exercise 2: some calculations

a. Use the elementary functions `/`, `-`, `^` and the function `sum` to calculate the mean

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

and the standard deviation

$$\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$$

of the fare paid by the titanic passengers.

b. Verify your answer by using the built-in R functions `mean` and `sd`.

RStudio

We will work with the dataset that gives the survival status of passengers on the Titanic. We provide the dataset in **Excel** format (Titanic3.xlsx). See this link for a description of the variables. Our data set has three additional columns: `dob` (date of birth), `family` (whether the person travelled with family) and `agecat` (age categorized in age ranges).

We assume that you have created a folder on your computer for the R course. Start the RStudio program from within that folder. Your working directory should be that folder. If it is not, set the working directory via

Session → Set Working Directory → Choose Directory...

Open a new R Script file via the menu:

File → New File → R Script...

Write the code in that R Script file.

Some useful Keyboard Shortcuts in RStudio for Windows

Over time it is useful to learn a couple of RStudio shortcuts that you frequently use. You can find an overview of them under the **Help → Keyboard Shortcuts help** menu. More shortcuts and tips can be found on the Appsilon website.

> **Note**
>
> On the mac, Ctrl → Cmd, Alt → Option, and Enter → Return.
>
> - **Ctrl+Enter**: Run current line
> - **Ctrl+Alt+I**: Insert new chunk
> - **Alt+-**: Insert assignment `<-`
> - **F1**: Show function help page when cursor is on function name
> - **Ctrl+Tab**: Switch between tabs in Source Editor
> - **Ctrl+1 / Ctrl+2**: Move focus to Source/Console
> - **Tab / Ctrl+Space**: Code suggestion/completion
> - **In Console**:
>   - Arrows **Up/Down** to navigate earlier commands
>   - **Ctrl+Up** pops up a listing of latest commands

Exercise 3: importing data from Excel and first check

Import the dataset by clicking on **Import Dataset** from the **Environment** pane. Browse for the Excel file. Assign the R object name `titanic` to the dataset under **Import Options**. Copy the code under **Code Preview**. Click on the "Import" button. Paste the code you have copied earlier to your **Script File**.

a. Inspect the data set in spreadsheet-like format that shows up after importing (if it does not show up, you can click on the `titanic` object in the **Environment** pane). Do the variables make sense? Are there any weird values? Sort the data in ascending order of age by clicking on the **age** in the data set.

b. Use the function `dim` to find the number of rows (passengers) and columns (variables) in the data set.

c. Use the `str` function to obtain a summary of the basic characteristics per variable. Have a look at the columns for `survived` and `age`. Do they have the mode that you expect? (You can also use the `mode` function to check their modes.)

Exercise 4: some columns summarized

A problem with the mode for age is that missing values are coded as a character value `"NA"` in our Excel file; it is not recognized as an R type NA value. In the future, you may encounter other representations of missing values, such as `missing`, `404`, `-99` or

`unknown`, which should also be treated as `NA`. Read in the data set as `titanic` object again, but specify `NA` as the missing value indicator under **Code Preview**. Your code under **Code Preview** should now look like

```
library(readxl)
titanic <- read_excel("data/raw/Titanic3.xlsx", na = "NA")
View(titanic)
```

a. Issue the commands

```
head(titanic, 10)
tail(titanic, 10)
```

What do you observe?

b. The `readxl` package that is used to import the data stores the data as a `tibble`, which is a special type of format on top of the `data.frame` format to store data in R. Override the `tibble` format using the function `as.data.frame`.

```
titanic <- as.data.frame(titanic)
```

Run the `head` and `tail` functions again. What has changed?

c. Summarize the whole dataset using the `summary` function. What do you observe for each of the variables?

d. Compute the 5%, 25%, 50%, 75% and 95% quantiles, the inter-quartile range and the standard deviation for age and fare. You may have to take a look at the `help` files for the functions `quantile`, `IQR` and `sd`.

e. For categorical variables, `summary` is not very informative. Use the `table` function to summarize the variables `pclass`, and `sex`. Also give a two-by-two table for sex and survival status using the same `table` function. Again, have a look at the help file if needed.

f. The function `addmargins` adds the row sums and the column sums to the table of survival status by sex. The function `proportions` tabulates proportions instead of absolute numbers.

```
addmargins(table(titanic$sex, titanic$survived))
proportions(table(titanic$sex, titanic$survived))
```

Use the `proportions` function to compute the fraction (or percentage) of survivors by sex. Have a look at the help file if needed.

Exercise 5: work with subsets

a. Take a look at the name, and the home town/destination of all passengers who were older than 70 years. Use the appropriate selection functions.

It is seen that there is one person from Uruguay in this group. Select the single record from that person. Did this person travel with relatives?

b. Make a table of survivor status by sex, but only for the first class passengers. What do you observe? Compare this with the survival status for the 3rd class passengers.

Exercise 6: make factor variables

a. Make passenger class and sex into factor variables.

b. Add a variable `status` to the Titanic data set, which gives the survival status of the passengers as a factor variable with labels "no" and "yes" describing whether individuals survived. Make the value "no" the first level.

c. Run the `summary` function again. What change do you observe compared to Exercise 4?

Exercise 7: working with dates

a. The variable `dob` gives the date of birth, using days since January 1st, 1960 (this is the time origin used by Stata). Transform this to a true date value using the function `as.Date` (specify the origin argument).

b. The default display is "year/month/day". Make a table of the day of the month that the passengers were born, using the `format` function. What do you observe?

c. What was the earliest date of birth among all the passengers on the boat. Does this correspond to the age of the oldest person on the date that the ship sank (April 15th, 1912)? Use the functions `min` and `max`.