Exercises for Lecture 3

Statistical Computing with R, 2023-24

Exercise 1: preparing your R script and working directory

1. If you haven't done so yet, create a folder in your laptop where you can save all the files that you will download and create during this course
2. Create a subfolder for the material from lecture 3, and save today's slides and material from the practical therein
3. Create an `R` script, and save it in the same subfolder that you created at point (2)
4. Write your solutions to the exercises below in the `R` script that you just created. Before doing that, set the working directory to the subfolder containing the material from today's class.
5. Use comments to separate the different exercises and questions, as well as to comment the code itself
6. Don't forget to save your progress regularly!

Exercise 2

Let $v = (v_1, ..., v_n)$ be a numeric vector of length n. The vector has mean $\bar{v} = \frac{1}{n}\sum_i v_i$ and sample variance $s_v^2 = \frac{1}{n-1}\sum_i (v_i - \bar{v})^2$. Consider the following operations:

- centering: $v_C = v - \bar{v}$;
- scaling: $v_S = \frac{v}{s_v}$;
- normalization: $v_N = \frac{v-\bar{v}}{s_v}$.

In this exercise, you will create functions that implement these 3 operations.

1. Write a function that computes v_C.
2. Write a function that computes v_S.
3. Write a function that computes v_N.

Exercise 3

Let `v = c(2, 4, 17, -8, -2, 3, 6:8, -5, -2)`.

1. Compute $\bar{v}$ and σ_v^2.

2. Compute v_C using the function that you wrote in exercise 1. What are its mean and variance?
3. Compute v_S. What are its mean and variance?
4. Compute v_N. What are its mean and variance?

Exercise 4

Write a function that given a numeric vector as input, returns as output a data frame with the following summary statistics: minimum, first quartile, median, mean, third quartile, maximum. The output data frame should contain two columns, one with the name of each statistics and the other with the corresponding values. In other words, it should look like this:

```
##          quantity value
## 1         minimum   0.7
## 2 first quartile   1.2
## 3          median   2.5
## 4            mean   2.4
## 5 third quartile   3.0
## 6         maximum   8.0
```

Apply the function to the two variables present in the `cars` data frame (see `?cars` and `View(cars)`).

Exercise 5

During the lecture we have seen how we can use the function `skewness( )` from the `moments` package to compute the skewness of a random variable. Now, it's time to write your own function to compute the skewness for a given input `x`!

Recall that the skewness index γ is defined as

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}, \text{ where } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

1. Write a function that given a vector `v` as input, computes its skewness (for now, ignore the possibility that `v` might contain missing values)
2. Generate 1000 random numbers from binomial distributions with the following parameters: $n = 3$ and $p = 0.1, 0.5, 0.8$. Compute the skewness associated to the different values of p.

Exercise 6

Let's consider again the `heights` and `wages` datasets from the `brolgar` package (in case you forgot about them: see the exercises from lecture 2).

1. Compute the skewness of the distribution of mean heights in the following years: 1900, 1950, 2000. Do it using both the function that you wrote to compute skewness, and the `skewness` function from the `moments` package. Do you obtain the same results?
2. Compute the skewness of hourly wages for workers with at most 1 year of work experience. Is the distribution left (negatively) or right (positively) skewed?

Exercise 7

At the URL https://data.worldbank.org/indicator/SP.POP.TOTL you can find data on the total population of all world countries. In the *Download* section you can obtain data in the following formats: csv (a folder containing 3 csv files), xml, and xls.

1. Download the csv and xlsx files, and store them in the subfolder that you created in exercise 1.

Two the 3 csv files you downloaded contain information that is relevant for us: one contains the population counts by country, and the other metadata about each country.

2. Open these two csv files with a viewer of your choice (e.g. NotePad++, Excel, OpenOffice Calc, Numbers...). Pay attention to how the data are organized. For example: do the data have headers? Do they start in the first row, or later?
3. Import the two files as datasets in `R`, and respectively call them `population` and `metadata`.
4. Save `population` to an `.RData` file.
5. Save `population` and `metadata` in the same `.RData` file.

Now, let's try to import the same data from xlsx. To do so, you need to install the `readxl` package first, and then load it:

```
install.packages("readxl")
library(readxl)
```

Differently from csv files, where 1 data table = 1 file, xls files can contain multiple data tables, organized as Sheets. In this case, the *Data* sheet contains the population data, whereas the *Metadata - Countries* one contains the metadata.

7. Use the `read_excel` function to import both the population data, and the metadata
8. Which country currently has the largest population (use information from the last available year to answer this question)? And which one the smallest?