Statistics with R – Exercise 2

Note: Use Rmarkdown for Exercise 2. Please updload a .PDF file including figures! In case you have problems with Rmarkdown please upload the .RMD file as well. All rules in the *Guidelines* document are still valid! Solve the tasks below using R and upload your solutions until Monday, the 14th of December (group registration until 11th of December).

Task 1 – Robustness of mean/median ($3 + 3 = 6$ points):

Some comments first: Usually we take only a single sample of size n from a population, e.g. we use the code

```r
set.seed(123)
x <- rnorm(n = 50, mean = 0, sd = 1)
```

to take a sample of size $n = 50$ from a standard normal distribution. Sometimes[1] we can take N samples, each of size n. In R this can be realized by taking a single sample of size $n \cdot N$ and rearranging this vector in a $n \times N$ matrix, in which each of the N columns represents a single sample of size n.

To estimate the expectation of our population/random variable, we used the **sample mean** $\overline{x}$. Unfortunately, the mean is not a robust measure, i.e. its value (and therefore the estimate) can be greatly influenced by a few outliers. For symmetric distributions (such as the normal distribution), we can also use the more robust **sample median** x_{med} as an estimate for the expectation. In this task we want to investigate the properties of these two estimates by drawing not a single, but many samples:

1. Draw $N = 1000$ samples, each of size $n = 50$, from a $\mathcal{N}(0, 1)$ and calculate the mean and median for each sample (this results in two vectors of length 1000). Visualize the estimates in two side-by-side boxplots (one for the mean estimates, one for the median estimates)[2]. What is the main difference between these two? Which estimation method (mean or median) would you choose, if you know that the underlying population is normally distributed, as it is the case here?

2. Now we simulate the situation of outlying observations in our data. Again, create $N = 1000$ samples, each of size $n = 50$, but now only 90 % (45 observations) of each sample shall come from a $\mathcal{N}(0, 1)$, whereas the remaining 5 observations shall be drawn from an exponential distribution with `rate = 0.2` (see the help page for `rexp()`). Create mean and median estimates and visualize the results in the same way as before. Which estimation method (mean oder median) would you choose, if you assume outliers in the data?

[1] In most cases the number of samples N and/or the sample size n are limited by costs and time. If we use a computer for taking a sample, we are only limited by memory and computer speed.

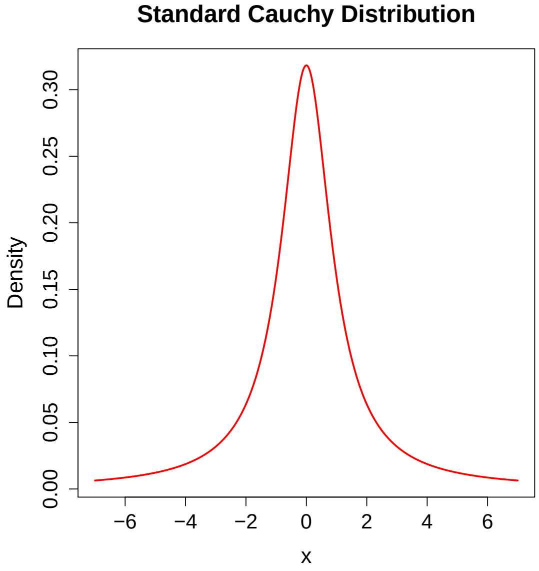
[2] For a better comparison of the two boxplots, you might find the argument `ylim = c(a,b)` useful, which sets the plotting range of the vertical axis to the interval $[a, b]$. Of course, you have to provide some numeric values for a and b.

Task 2 – Estimation ($1 + 1 + 1 = 3$ points):

The so-called **Standard Cauchy distribution** has the density

$$f(x) = \frac{1}{\pi \cdot (1 + x^2)} \qquad \text{for } x \in (-\infty; +\infty)$$

and looks somewhat similar to the Standard Normal distribution $\mathcal{N}(0, 1)$:



In Task 4 of the first exercise we learnt that we can estimate population parameters (such as the expectation or variance) with a higher precision, if we have a larger sample. Now we want to see, if this also works for the Cauchy distribution:

1. Draw samples of size $n_1 = 100$, $n_2 = 5000$ and $n_3 = 100000$ from a Standard Cauchy distribution and estimate the population mean and variance. You might need one of the functions `dchauchy()`, `rcauchy()`, `qcauchy()` and `pcauchy()`.

2. What do you observe? What could be the explanation?[3]

3. The main difference between the standard normal and standard cauchy distribution can easily be seen in two side-by-side boxplots. Take two random samples of size $n = 100$ – one from each distribution – and do the following:

```
par(mfrow = c(1,2))
# ... function for normal boxplot
# ... function for cauchy boxplot
par(mfrow = c(1,1))
```

   What is the most obvious difference between the two distributions/boxplots?

[3] A hint: Wikipedia is a good reference.

Task 3 – Working with a real data set $(2 + 1 + 1 + 1 + 2 + 1 + 1 + 2 + 1 + 1 + 1 + 2 = 16$ points):

In the *exercise 2* section on *Boku Learn* you will find a data file (`CO2`) in 4 different formats (`.RData`, `.xlsx`, `.txt` and `.csv`). All files contain the same data from Eurostat on ”Greenhouse gas emissions by source sector”(Metadata). File *exercise_2_help.pdf* provides information about this dataset. The dataset has five dimensions:

- Air pollutant (`airpol`): Greenhouse gases CO2, N2O, CH4, HFC, PFC, SF6, NF3 and their aggregate (GHG). The fluorinated gases and the GHG aggregate are in CO2 equivalents .
- Geopolitical entity (`geo`): EU Member States, EFTA Countries, Candidate Countries. See details under point 3.7
- Source sector for air emissions (`airemsect`): Sectors are classified according to the Common Reporting Format (CRF) in line with the UNFCCC reporting requirements. See details under point 3.3
- Period of time (`time`): Data are annual.
- Unit (`unit`): Thousand tonnes and million tonnes.

1. Use appropriate functions to import the data from all 4 files. For each file, convince yourself that the data were imported correctly (by viewing the first few rows of the data). It is of course sufficient, if all subsequent tasks are performed with one of these data files. After this task you can remove unnecessary objects with `rm()` function and use garbage collection (`gc()` function) to return memory to the operating system. If might be useful to use `gc()` after a large object (not a vector of size 10000) has been removed. **Example**:

```
rm(list = "aa") # remove object aa from Environment
rm(list = c("aa", "bb", "cc")) # remove several objects from Environment
gc()
```

2. Set seed to `as.numeric(format(Sys.time(), "%H%M%S"))`, create a vector containing the student id numbers from your group (omit leading zeros), using `sample()` function choose one id and according to it set random seed again. **Example**:

```
set.seed(as.numeric(format(Sys.time(), "%H%M%S")))
ids <- c(241235246, 112414, 12135236)
id <- sample(ids, 1)
set.seed(id)
```

3. Using functions `str()` and `summary()` take a look at the data structure. What is the type of each variable/column? How many observations (rows) and how many variables (columns) do we have? Are there any missing values? Which column has missing values and how many? Please answer in complete sentences.

4. Create a vector containing unique values of variable (column) `geo`. Using this vector randomly select 2 countries and assign them to an object named `geo_c`.

5. Filter out rows from the data (object from step 1), where:

   - `unit` is equal to `"THS_T"`(Thousand tonnes),
   - `airpol` is equal to `"GHG"` (Greenhouse gases),
   - `airemsect` values are %**in**% vector `c("CRF3", "CRF31", "CRF1A3")` (”Agriculture”, ”Livestock”, ”Fuel combustion in transport”),
   - `geo` values are %**in**% vector `geo_c`,
   - assign the result to a new object (the name can be chosen freely).

   Note that this is a chain of ”AND” statements and the result is a single object (not 4 different objects). You can also remove the original object and free up memory (functions `rm` and `gc`).

6. Remove columns/variables `unit` and ‘airpol‘, as they do not hold any information.

7. How many observations does each country (column `geo`) have?

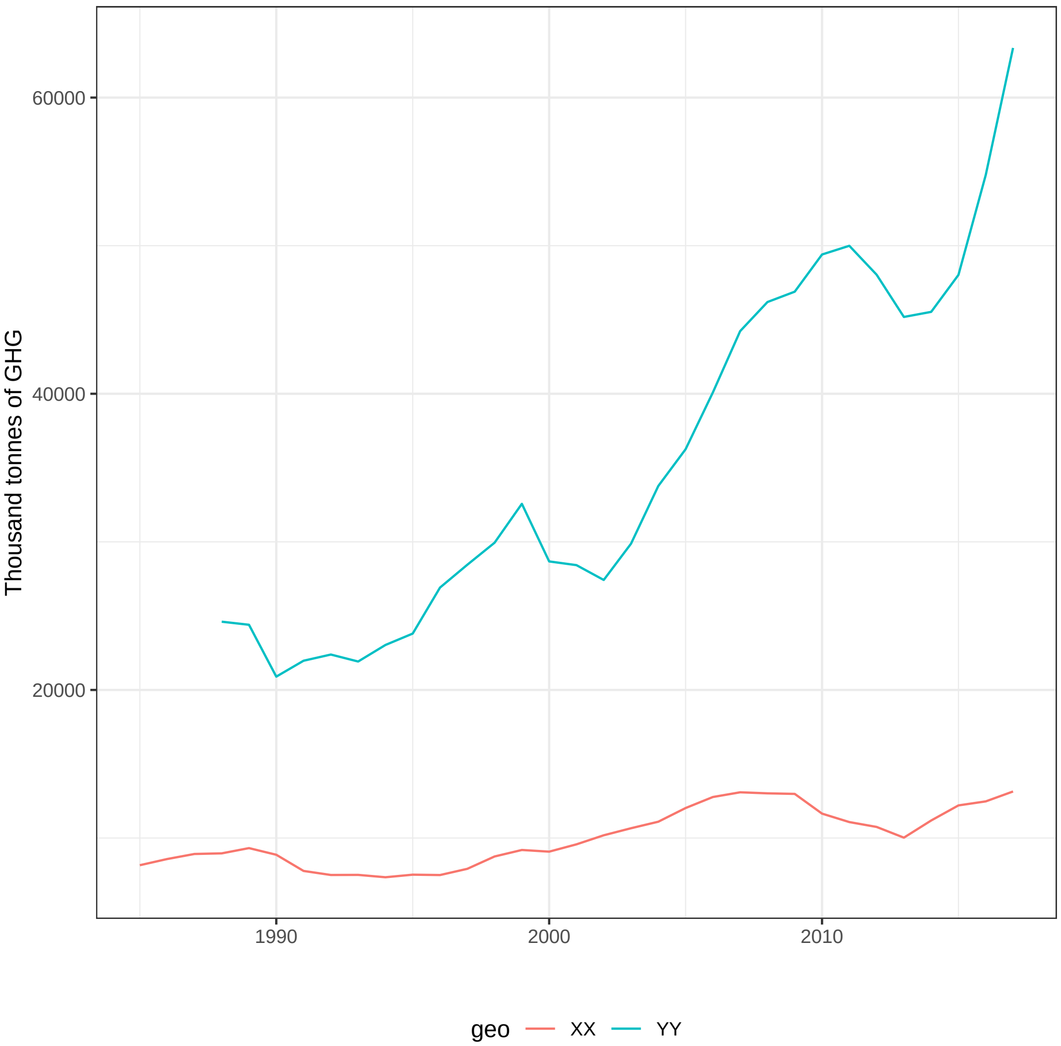
8. Labels used by Eurostat are quite hard to understand. Thus, change values in column `airemsect`: ”CRF1A3” into ”Transport”, ”CRF3” into ”Agriculture”, ”CRF31” into ”Livestock”. Also, rename `airemsect` into ‘Sector‘.

9. Calculate the average greenhouse gas emission in each sector. Which sector on average produced the most greenhouse gas emissions?

10. Calculate the average greenhouse gas(GHG) emissions in different sectors for each country. Which country has a higher average greenhouse gas emissions of ”Livestock”? By how much?

11. How much GHG did sector `"Livestock"` produce in each country for the period 2000-2017?

12. Make a plot (similar to the one below) with two separate lines (for each country) showing the values of GHG emissions of ”Transportation” sector through years.


60000
40000
20000
Thousand tonnes of GHG
1990
2000
2010
geo
XX
YY