Statistical Methods Fall 2017

Assignment 1: Exploring and summarising data

Deadline: November 8, 23.59h

Topics of this assignment

The exercises below concern the topics that were covered in Lecture 1 and in the beginning of Lecture 2: data and summarising data. Before making the assignment, study these topics. Numbers of exercises in the book refer to the twelfth edition (New Pearson International Edition).

How to do the exercises?

- Do the exercises as efficiently as possible. Some exercises or sub-parts of exercises do not need the use of *R*, and some do. Write your report in English. To hand in: create a single pdf file of your work including your name and group number. Upload on Canvas.
- Data files and/or local *R*-functions, that are needed for the assignment, are available on Canvas.
- The text of the report should not exceed 4 pages, this is excluding figures and the appendix with *R*-code.
- It is important to make clear in your answers how you have solved the questions: do not only give answers and results, but also motivate your answers. Put the relevant *R*-code ***in an appendix***. Do not copy *R*-code in the answers themselves, and only include in the appendix the code that led to your answer. Do not put entire data sets in the appendix.
- Graphs should be made and viewed on screen first; put the final version in your report. Multiple graphs can be put into one figure using the command `par(mfrow=c(k,r))`, see `help(par)`. Make sure the dimensions of the graphs are adequate and that figures are concise: one figure should not take up a whole page.
- In your report, round the results that you obtained from *R* to a suitable number of digits.

Not adhering to these rules may have as a consequence that some of your points will be deducted!

Theoretical exercises

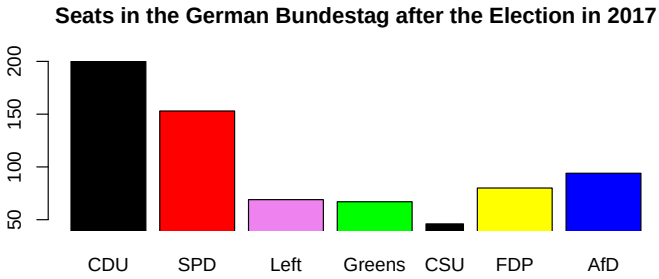
Exercise 1.1
Do exercises 10, 12, 26 of Section 1.2.

Exercise 1.2
Do exercises 22, 32 of Section 1.3.

Exercise 1.3
Do exercises 6, 12, 18 of Section 1.4.

Exercise 1.4

a. The following graph shows the distribution of seats in the German Bundestag. What is wrong with the presentation?



(Source: https://www.bundeswahlleiter.de/info/presse/mitteilungen/bundestagswahl-2017/32_17_vorlaeufiges_ergebnis.html)

Remark: It is no problem that "CDU" and "CSU" have the same color (black) because together, they form the "Union".

b. Suppose that you are preparing the annual report of a big social network company. One of your datasets contains the average numbers of daily public posts for each registered user. Which of the following graphs would be best for describing the distribution of the average number of posts: histogram; bar chart; Pareto chart; pie chart?
Which graph is the best to compare the means of all average numbers of posts in the subgroups "male, single", "female, single", "male, married", "female, married"?

R-exercises

Hints concerning R:

- For the exercises below you can use, for instance, the *R*-functions `hist`, `boxplot`, `mean`, `median`, `sd`, `min`, `max`, and `summary`. If necessary, experiment with the different options these functions have.
- The *R*-function `quantile(x,`α`)` gives the α-quantile of the values in the vector `x`. For example, `quantile(x,0.25)` gives the first quartile of x. Instead of one single value, also a vector $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ can be inserted for the parameter α in `quantile`. Check which output this function gives when the parameter α is not specified.

Exercise 1.5

a. Make a suitable histogram and boxplot for the data in the file `sampleA`.

b. Give one or more suitable numerical summaries for the location and the spread of the distribution of these data.

c. Based on your summaries in parts a and b, briefly answer for this data set as many of the basic questions (location, spread/variation, range, extremes, accumulations, symmetry,...) about the data distribution as possible.

d. Perform parts a, b and c for the data in the file `sampleB`.

e. Based on all results of parts a–d, do you think that the two data sets originate from the same population distribution? Why (not)?

Exercise 1.6 In the file `mileage` you can find data about fuel usage of cars. The first two components in the list give the fuel usage in miles per gallon and the number of cylinders of cars of type 1. The third and fourth component give the same quantities for cars of type 2. Look at the data first. Make an appropriate summary of these data, both graphically and numerically. Comment on these summaries. Is one type more fuel efficient than the other? Is it without any risk to directly compare the data of both types of cars?