Statistical Methods Fall 2021

Assignment 1: Exploring and summarising data

Deadline: see Canvas

Topics of this assignment

The exercises below concern the topics covered in Lecture 1 and at the beginning of Lecture 2: data and summarising data. Before getting started with the assignment, study these topics. Numbers of exercises in the book refer to the 12th edition (New Pearson International Edition).

How to do the exercises? (Also take a look at Assignment_Example.pdf on Canvas!)

- Solve the exercises as efficiently as possible. Some exercises or their sub-questions of exercises do not require you to use R, while some others do. Write your report in English. To hand in: create a single PDF file of your work including your name and group number and upload it on Canvas.
- Data files and/or local R-functions 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, executable, and copiable** R-code (without the prompt sign ">") ***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: a single 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 *In a)-c), name the chosen sampling method and determine whether the sampling method seems to be sound or is flawed. Always motivate your choice.*

a) In a survey on COVID-19 vaccinations, the Dutch Central Bureau of Statistics randomly selected and mailed 2052 teens (aged 12-17) about their vaccination status.

b) In another survey on COVID-19 vaccinations, the Dutch Central Bureau of Statistics randomly selected 20 secondary schools in The Netherlands and asked all of their students about their vaccination status.

Also, explain what is wrong in c).

c) In an online poll conducted by a big Dutch online newspaper, 5012 Internet users chose to respond, and 76% of them stated that they were fully vaccinated against COVID-19.

Exercise 1.2 *Determine which of the four levels of measurement (nominal, ordinal, interval, ratio) is most appropriate. Also, if there is anything wrong with the given summary statistics, explain what is wrong.*

a) A survey about the public transport system should measure the agreement to the statement "I enjoy taking off-peak Intercity trains." The survey used a 5-point Likert scale with the following options and numbers of replies: "Strongly disagree" (11 times), "Disagree" (7), "Neither agree nor disagree" (13), "Agree" (20), "Strongly agree" (37). The statisticians who conducted the study reported that the mean was slightly below "Agree".

b) A Dutch bank analyzed the balance of 1000 randomly selected bank accounts of their customers. The first 5 balances (in Euros) were: 1,167.51, 2,614.12, 4,698.06, -152.63, 307.95. The average (mean) balance of all 1000 accounts was 1,714.42 Euros with a standard deviation of 625,63 Euros.

Exercise 1.3 *In a), determine whether the given description corresponds to an observational study or an experiment. Give a brief explanation of your choice.*

a) In a clinical trial of the cholesterol drug Lipitor, 190 subjects were given 21-mg doses of the drug, and 3.8% of them experienced nausea.

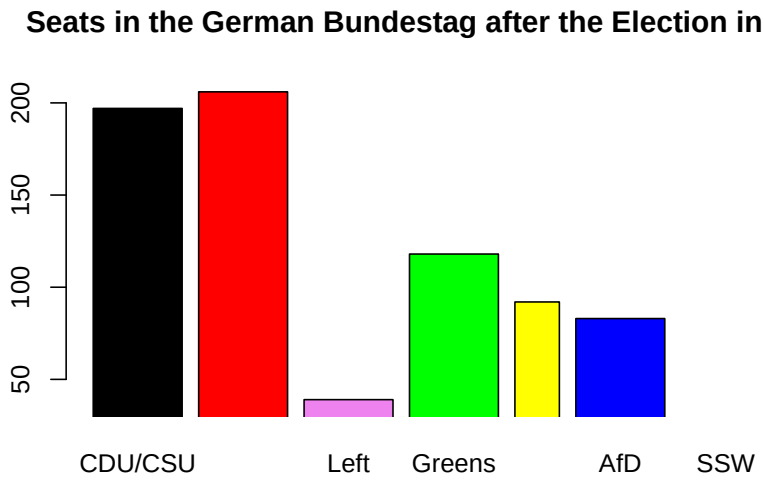
In b) and c), identify which of these types of sampling is used: random, systematic, convenience, stratified, or cluster.

b) When collecting data from different sample locations in a lake, a researcher uses the "line transect method" by stretching a rope across the lake and collecting samples at every interval of 10 meters.

c) On the day of the last presidental election, a television channel organized an exit poll in which specific polling stations were randomly selected and all voters were surveyed as they left the premises.

Exercise 1.4

a) The graph below shows the seat distribution in the 20-th German Bundestag (after the election in September 2021). The seat distribution is as follows: CDU/CSU: 197, SPD: 206, Left: 39, Greens: 118, FDP: 92, AfD: 83, SSW: 1. What is wrong with the presentation?



b) Human resources departments of several IT companies were surveyed about areas in which job applicants make mistakes (multiple choices possible). The areas found in the survey were: interview, résumé, cover letter, reference checks, interview follow-up. Which of the following graphs would be best for describing the mistakes: histogram; bar chart; Pareto chart; pie chart?

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 *Always describe the most important findings in numerical and graphical summaries.*

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 or 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 appropriate graphical and numerical summaries of those cars of type 1 which have 4 cylinders. Repeat the same for the cars of type 2 which have 4 cylinders.
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 when including also the cars with more cylinders?