SDA 2022 — Assignment 5

For these exercises you can make use of the following R-functions:
for 5.1, `binom.test` for the sign[1] and the binomial test;
for 5.2, `mad` to compute the median absolute deviation, i.e. the MAD estimator for spread; `binom.test` for the sign test[1], `wilcox.test` for the signed rank test, `t.test` for the *t*-test;
for 5.3, `wilcox.test` for the Mann–Whitney test and `ks.test` for the two sample Kolmogorov–Smirnov test.
Additional tests might be applied as well; some of them have to be done manually!
You may also again use the function `bootstrap` contained in the file "functions_Ch5.txt".

Make a concise report of *all* your answers in *one single PDF file*, with only *relevant R* code *in an appendix.* It is important to make clear in your answers how you have solved the questions. Graphs should look neat (label the axes, give titles, use correct dimensions etc.). Multiple graphs can be put into one figure using the command `par(mfrow=c(k,l))`, see `help(par)`. Sometimes there might be additional information on what exactly has to be handed in. **Read the file AssignmentFormat.pdf on Canvas carefully**.

General information on the .RData file to be submitted to the dummy assignment "RData A5" on Canvas:
Create in R a list `mylist` that contains the required entries as specified in the exercise(s) which are marked as "(partially) .RData file hand-in". You should store your list in an .RData file by using the R command `save(mylist, file="[PATH]/myfile5_[GROUP_NO].RData")`, where `[PATH]` stands for the path on your computer where you wish to save the .RData file and `[GROUP_NO]` stands for the number of the assignment group you have chosen on Canvas). For example, for if you're in group 91, use `"[PATH]/myfile5_91.RData"`.

Note: This is *one* list in *one* file for all exercises of this assignment together.

In any case, `mylist` must have the entry `stud_no`: a vector that contains the student numbers of you and your group partner.

For your convenience, you could use the following skeleton R command to create the list; the entries still need to be specified and some more must be added (instead of the "..."):

```
mylist <- list(stud_no = c(... , ...),
               "1_a_pval" = ...,
               "1_a_reject" = ... (TRUE or FALSE),
               "1_b_pval" = ...,
               ...
               "2_b_sd" = ...,
               "2_d_mad" = ...,
               "2_h_CI_sign" = c(... , ...),
               ...)
```

[1] See Lecture 7 for details on how to do this!

Exercise 5.1 (.RData file hand-in) This exercise is about the grades for one of the resit exams for a statistics course for the Computer Science students. The data can be found in the file `statgrades.txt`. We assume that these data are a random sample that is representative for the grades from resit exams in any other year the statistic course took place. Denote by m the unknown median of the unknown true grades distribution.

a. Test $H_0 : m \geq 6.2$ against $H_a : m < 6.2$ at level $\alpha = 1\%$.

b. Test $H_0 : m = 6$ against $H_a : m \neq 6$ at level $\alpha = 5\%$.

c. Denote by p the probability to get a grade of at least 5.5.
Test $H_0 : p \leq 45\%$ against $H_a : p > 45\%$ at level $\alpha = 10\%$.

Hand in (all stored in your .RData file): the following entries of your list `mylist` in R:

a.: `"1_a_pval"`: p-value of the test in a. i.e. a single number between 0 and 1,
`"1_a_reject"`: boolean indicating rejection of H_0 i.e. `TRUE or FALSE` (not as a string),

b.: `"1_b_pval"`: p-value of the test in b. i.e. a single number between 0 and 1,
`"1_b_reject"`: boolean indicating rejection of H_0 i.e. `TRUE or FALSE` (not as a string),

c.: `"1_c_pval"`: p-value of the test in c. i.e. a single number between 0 and 1,
`"1_c_reject"`: boolean indicating rejection of H_0 i.e. `TRUE or FALSE` (not as a string).

Exercise 5.2 (partially .RData file hand-in) With *cloud seeding* a small airplane is used to add a particular substance to clouds in order to change the precipitation properties[2]. In a cloud seeding experiment in 1975, precipitation values of two groups of clouds, *seeded* and *unseeded*, were compared. The precipitation data of this experiment are contained in the file `clouds.txt`. In this exercise, we only focus on the data set `unseeded`.

a. Investigate the data graphically and numerically.[3]

b. Determine the sample standard deviation of the data set; this is often used as a measure for the accuracy of the measurements.

c. Determine a bootstrap estimate of the standard deviation of the estimator of accuracy used in part b.[4]

d. Repeat parts b and c, but now use as a measure for the accuracy of the measurements the MAD which is a more robust estimator for spread.[4]

e. Which estimator for the accuracy do you prefer for these data, the sample standard deviation or the MAD? Explain how you reached your conclusion.

f. Which test do you prefer for testing the location of the precipitation values of the unseeded clouds: the t-test, the sign test, or the signed rank test? Motivate your answer.

g. Test whether the location of the precipitation value of the unseeded clouds is less than 40 by using the test your preferred in part f. Take the significance level $\alpha = 0.05$.

h. Make two-sided 95% confidence intervals for the location of the precipitation value of unseeded clouds based on the sign test, the signed rank test, and the t-test.
Note: Warning messages from the `wilcox.test` *command may be ignored.*

i. Which confidence interval from h. do you value most, and why?

Hand in:
In your main report: Results of parts a, c, d, and g, and your answers to parts e, f, and i.
Stored in your .RData file:

b.: `"2_b_sd"`: the sample standard deviation based on the `unseeded` dataset i.e. a single positive number,

d.: `"2_d_mad"`: the median absoulte deviation based on the `unseeded` dataset i.e. a single positive number,

h.: `"2_h_CI_sign"`: confidence interval for the location parameter based on the sign test i.e. an interval specified by a vector consisting of the lower and upper boundary,

h.: `"2_h_CI_wilcox"`: confidence interval for the location parameter based on the signed rank test i.e. an interval specified by a vector consisting of the lower and upper boundary,

h.: `"2_h_CI_t"`: confidence interval for the location parameter based on the t-test i.e. an interval specified by a vector consisting of the lower and upper boundary.

[2] For more on that see, e.g. `http://en.wikipedia.org/wiki/Cloud_seeding`.

[3] The gained insights might turn out handy in one of the later parts of this exercise.

[4] Use a sufficiently large number of bootstrap replicates to make sure that the estimate of the standard deviation is not heavily influenced by randomness.

Exercise 5.3 In 1882 S. Newcomb performed a series of experiments to determine the speed of light with the method of Foucault. Light bounced from a fast rotating mirror in Fort Meyer on the west bank of the Potomac in Washington to a fixed mirror at the foot of Washington monument, and back to the rotating mirror. The speed of the light was calculated from the measured distance between the mirrors (3721 meter) and the deflection angle of the emitted and received light on the rotating mirror. The file `newcomb.txt` contains Newcomb's data; *the given values times* 10^{-3} *plus* 24.8 *are the original observations of the time (in sec*$^{-6}$*) the light needed for traveling twice the 3721 meters.*

a. Investigate whether there is a difference between the first 20 and the last 46 observations. Do this by making suitable graphical summaries, by determining an estimate of the difference, and by performing one or more suitable tests at level $\alpha = 5\%$.
   Explanation: 1. The graphical summaries should visualize the datasets and make them comparable. Possibly choose the same scales for the axes.
   2. Choose a suitable estimator and a suitable test. Motivate your choice in one or two sentences.

b. Determine a 95% confidence interval for the traveling time of the light in Newcomb's experiment. According to present day physics the true value of the traveling time is equal to 24.8332 (in sec^{-6}). Is this value in the computed confidence interval? What is your conclusion? *Explanation: choose a suitable location estimator. Motivate your choice in one or two sentences.*
   Note: to compare the given true travling time to the confidence intervals based on the data, you should subtract 24.8 from the true time and then multiply the result with 1000.

Hand in: relevant plots and numbers and your answers to the questions.