SDA 2019 — Assignment 6

To compute the α-trimmed mean of data `x` use the *R*-command `mean(x,trim=`α`)`. Bootstrap values for trimmed means can be computed by `bootstrap(x,mean,B,trim=`α`)`.

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**.

Exercise 6.1 In 1879 and 1882 Michelson performed experiments to determine the speed of light. The measurements minus 299000 are given in the file `light.txt`[1]. For the composite null hypothesis *"H_0: X is normally distributed"* the standard Kolmogorov-Smirnov test cannot be used. To test this null hypothesis an adjusted Kolmogorov-Smirnov statistic can be used, which does not have the distribution (and corresponding p-values) of the standard Kolmogorov-Smirnov test statistic D_n. Its distribution under the null hypothesis and the corresponding p-values can be estimated by means of the bootstrap method.

a. In the lecture, we have mathematically shown that the Kolmogorov-Smirnov test statistic is distribution-free. Discuss whether the arguments used in that proof can also be used to show the same for the adjusted Kolmogorov-Smirnov statistic

$$\tilde{D}_n = \sup_x |\hat{F}_n(x) - \Phi((x - \bar{X})/S)|.$$

In addition, explain why $\tilde{D}_n$ is independent of the location and scale parameters of the data.

Note 1: In $\tilde{D}_n$ the empirical distribution function $\hat{F}_n$ is compared to the normal distribution with expectation $\bar{X}$ and variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. (cf. Figure 3.11 in the syllabus).

b. Test the composite null hypotheses that the measurement errors in 1879 and 1882 have a normal distribution using the adjusted Kolmogorov-Smirnov test statistic $\tilde{D}_n$.

Note 2: Values of D_n or $\tilde{D}_n$ can be extracted from the output of `ks.test` using `ks.test(data,dist,par)$statistic` with appropriate arguments `dist` (distribution) and `par` (parameters). Ignore warning messages of *R* about ties for this exercise.

Note 3: The bootstrap samples should follow the null hypothesis, so think carefully about how these should be generated. Next, compute for each of the bootstrap samples the value of the test statistic $\tilde{D}_n$. Finally, based on these bootstrap values the test can be performed by determining the p-value of the (right-tailed) test.

c. Are the two p-values found in part b different from the ones that you would have found with the standard Kolmogorov-Smirnov test (in the output of `ks.test` with the estimated mean and standard deviation)? If yes, explain the reason for the difference.

[1]For importing the data use the command `source("light.txt")`. In your global environment you will then find a list called "light".

Hand in: Your answers to part a, results of part b, and your answer to part c.

Background: *Michelson used the method of the French physicist Foucault. Light bounces from a fast rotating mirror to a fixed mirror at a distance and back to the rotating mirror. The speed of the light is calculated from the measured distance between the mirrors and the deflection angle of the emitted and received light on the rotating mirror. In the first series of experiments the distance between the mirrors was 600 m and in the second series 3721 m. Theoretically the observations in the second series should be 24 smaller than those in the first series.*

Exercise 6.2 First, study Example 5.3 in the syllabus.
The data in this example are contained in the file `lepton.txt`[2].

a. Explain in your own words what a 'weighted mean' is.
NB. You do not need to reproduce the outcomes!

b. Which mathematical calculations were done to find the confidence interval $[15.41, 21.09]$ that was obtained in the syllabus?

c. Make boxplots of the original data and briefly comment on what the boxplots show about these data.

d. Compute a bootstrap confidence interval for the percentage of D particles based on the 20% trimmed (unweighted) means.

Hint: for finding bootstrap values of D, you have to bootstrap the data of the different particles separately, as in Example 4.4 in the syllabus. In that example medians (50% trimmed means) are used, whereas here you are asked to use 20% trimmed means.

e. Do the same, but now based on the usual (unweighted, untrimmed) means.

f. Which conclusion do you draw about the existence of the unknown particle?

Hand in: your answers to parts a, b, and f, plots of part c, and results of parts d and e.

[2]For importing the data use the command `source("lepton.txt")`.