SDA 2019 — Assignment 5

For these exercises you can use the function `bootstrap` on the Canvas page (see Assignment 4). The R-function `quantile(x,`α`)` gives the α-quantile of values in the vector `x`. For the parameter α, either a single value or a vector $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ can be inserted into the function `quantile`.

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 <u>how</u> 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**.

Note: If it is not specified which bootstrap estimator to use, take the empirical and not the parametric one.

Exercise 5.1 Read Examples 3.4 and 4.4 in the syllabus about data on β-thromboglobulin levels which can be loaded by the R-code in `thromboglobulin.txt`[1]. You can select e.g. the `PRRP` data using R-command `thromboglobulin$PRRP` or `thromboglobulin[[1]]`. Or use `attach(thromboglobulin)` so that the variables `PRRP`, `SDRP` and `CTRP` are defined, see `help(attach)`.

a. Determine a 95%-bootstrap confidence interval for the expectation of the underlying distribution of `PRRP`. Take B sufficiently large.

b. Determine a 95%-bootstrap confidence interval for the median of the underlying distribution of `PRRP`. Take B sufficiently large.

c. Compare the answers of parts a and b. Which estimator of location do you prefer and why?

d. Determine a 95%-bootstrap confidence interval for the difference in mean between the two groups `SDRP` and `PRRP`. What can you conclude from this interval about the difference in mean of the two underlying distributions? (Note that this is a two sample problem, like in Example 4.4.)

Hand in: the computed intervals and your answers to parts c and d.

[1]For importing the data use the command `source("thromboglobulin.txt")`.

Exercise 5.2

This exercise illustrates an example where the bootstrap does not work very well. Before you make this exercise, first read Section 4.5 of the syllabus.

Consider $X_1, \ldots, X_n$, a sample from the uniform distribution on $[0, \theta]$ with $\theta > 0$ unknown. The estimator $T_n = \frac{n+1}{n} X_{(n)}$ is an unbiased estimator of the unknown θ. Note that $X_{(n)} = \max_i X_i$.

a. Generate a sample of size 50 from the uniform distribution on [0,1]. Compute, using the empirical bootstrap method, an estimate of the variance of T_n. Take $B = 1000$ bootstrap samples.

b. Repeat the whole procedure in part a. a few times (taking a new initial sample of size 50 each time) and compare the obtained estimates with the theoretical value for the variance of T_n (see e.g. the syllabus of the lecture Statistics).

c. Explain how the *parametric* bootstrap method can be used in this case. Perform this parametric bootstrap procedure to obtain again an estimate for the variance of T_n. Take $B = 1000$ bootstrap samples. Also repeat this *whole* procedure a few times.

d. Which of the two bootstrap methods, the empirical or the parametric, works better in this case? Can you explain why?

Hand in: your results of parts a, b, and c, and your answer to part d.