SDA 2019 — Assignment 4

For these exercises you can use the function `bootstrap` on the Canvas page (contained in the file "functions_Ch4.txt". Investigate this function before using it.

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 4.1 One sample drawn from a Gamma distribution with unknown shape and scale parameters, $k > 0$ and $\theta > 0$, respectively, is stored in the file `gamma.txt`. With the help of this sample, we would like to estimate the distribution of the sample skewess statistic.

Consider the method of moments estimators[1] $\hat{k} = \bar{X}_n^2/\hat{\sigma}_n^2$ and $\hat{\theta} = \hat{\sigma}_n^2/\bar{X}_n$, where $\bar{X}_n$ and $\hat{\sigma}_n^2$ are the sample mean and variance, respectively.

a. Write a function `skewness(x)` that returns the skewness of a sample `x`. Use it to estimate the skewness of the sample.
NB. if you found an R package that offers such a function, you may use it instead of programming it from scratch. In this case, report which library you used and what the function's name is; this should also be apparent from your R code in the appendix.

b. Use the empirical bootstrap method applied to our gamma sample to generate $B = 1000$ bootstrap estimates of the skewness statistic. Store these in a vector `empBS` in your R environment.

c. Repeat the steps of part b. but with the parametric bootstrap instead of the empirical bootstrap. Use the method of moment parameter estimates $\hat{k}$ and $\hat{\theta}$ as described above. Call the vector which containes the obtained bootstrap statistics `parBS`.

d. Plot two separate histograms of the bootstrap samples obtained in c. and d.
Compare the histograms with the skewness estimate of part a.
Now compare the histograms with the true value of the skewness which underlied the data generating process of our sample: 1.265.
Based on these comparisons, which bootstrap method seems preferable in the present context. Motivate your answer?

e. Use the empirical and the parametric bootstrap samples to find estimates of the variance of the skewness statistic.

Hand in: answers to the questions and relevant plots.

[1] The method of moment estimators are found by solving the first two moment equations $E(X) = k\theta \stackrel{!}{=} \bar{X}_n$ and $var(X) = k\theta^2 \stackrel{!}{=} \hat{\sigma}_n^2$ for k and θ. Here, X has a Gamma distribution with parameters k and θ.

Exercise 4.2 The file `birthweight.txt` contains data on birth weights (in grams).

a. Explore the distribution of the birth weight data graphically and find an appropriate distribution from which the data could have originated.

b. Estimate the median birth weight and find a bootstrap estimate of its variance. Explain which bootstrap method you chose and motivate this choice.

c. Now, repeat part b. but use as a bootstrap method a parametric bootstrap based on an exponential distribution with suitably estimated rate parameter.[2] Compare the resulting variance estimate of the median statistic with the one found in part b. Explain, with reference to the theory of Lecture 4, what went wrong.

d. Reprogram part c. so that everything, the variance calculation, the (e.g. $B = 1000$) bootstrap samples generation, and the parameter estimation is done in exactly one line in R code. Avoid the use of loops.
NB. You could, for example, use a clever combination of the R functions `var`, `replicate`, `median`, `rexp`, `mean`.

Hand in: relevant plots, results and answers to the questions, and your comments.

[2]For an exponential distribution, the rate parameter is one over the mean.