SDA 2021 — Assignment 4

For these exercises you can use the function `bootstrap` on the Canvas page (contained in the file "functions_Ch5.txt". Investigate this function before using it.

The *R*-function `quantile(x,`α) gives the α-quantile of values in the vector `x`. For the parameter α, either use a single value or a vector $(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

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 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 where the data could have originated from.

b. Estimate the median birth weight and find a bootstrap estimate of the standard deviation of the sample median. Motivate and explain your choice of bootstrap method.

c. Now, repeat part b but use as a bootstrap method a parametric bootstrap based on an exponential distribution with suitably estimated rate parameter.[1] Compare the resulting estimate of the standard deviation of the sample median statistic with the one found in part b. Refer to the theory of Lecture 5, to explain what went wrong.

d. Reprogram part c so that everything, the standard deviation 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.

Exercise 4.2 Read Examples 3.4 and 5.4 in the syllabus about data on β-thromboglobulin levels which can be loaded by the *R*-code in `thromboglobulin.txt`[2]. You can select e.g. the `PRRP` data using *R*-command `thromboglobulin$PRRP` or `thromboglobulin[[1]]`. Or use `attach(thromboglobulin)`[3] so that the variables `PRRP`, `SDRP` and `CTRP` are defined.

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

b. Repeat part a with the median instead of the mean.

c. Compare the answers of 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: this is a *two sample problem*, like in Example 5.4.

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

[1] For an exponential distribution, the rate parameter is one over the mean.
[2] For importing the data use the command `source("thromboglobulin.txt")`.
[3] see `help(attach)`

Exercise 4.3 In 1879 and 1882 Michelson performed experiments to determine the speed of light. The measurements minus 299000 are given in the file `light.txt`[4]. 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. The adjusted Kolmogorov-Smirnov statistic is

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

Its distribution under the null hypothesis and the corresponding p-values can be estimated by means of the bootstrap method.

a. Explain why $\tilde{D}_n$ is independent of the location and scale parameters of the data.

   Hint: The supremum is attained for some $x = X_i$ or immediately to the left of some X_i. Write $X_i \sim N(a, b^2)$ as $X_i = a + bY_i$ for $Y_i \sim N(0, 1)$.

   Note 1: In $\tilde{D}_n$ the empirical distribution function $\hat{F}_n$ is compared to the normal distribution with mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.
   (Cf. Figure 3.11 in the syllabus for a comparison of a theoretical cumulative distribution function to an empirical cumulative distribution function).

b. Test the composite null hypotheses that the measurement errors in 1879 and 1882 have a normal distribution; use 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; perhaps the procedure in Example 5.5 in the Syllabus helps you to get an idea. Next, compute for the i-th bootstrap sample the value of the bootstrapped test statistic $\tilde{D}^*_{n;i}$, $i = 1, \ldots, B$ (e.g. $B = 1000$). Finally, use these bootstrapped values of the test statistic to find a suitable p-value of the (right-tailed) test.

c. Are the two p-values found in part b different from those one obtains from the output of `ks.test`, when one uses as an input for the latter (for `par`) the estimated mean and standard deviation?
   If yes, explain the reason for the difference.

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.*

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