

## SDA 2023 — 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,  $\alpha$ )` gives the  $\alpha$ -quantile of values in the vector `x`. For the parameter  $\alpha$ , either use a single value or a vector  $(\alpha_1, \alpha_2, \dots, \alpha_k)$ .

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,1))`, 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.**

For each exercise, you should set a seed (of your choosing). Please include your uses of the `set.seed` command in R in the R code that you put in your appendix so that all of your outcomes will be reproducible by the teaching assistants.

**Exercise 4.1** The file `birthweight.txt` contains data on birth weights (in grams).

- Explore the distribution of the birth weight data graphically and find an appropriate distribution where the data could have originated from.
- Estimate the 5%-quantile of birth weights and find a bootstrap estimate of the standard deviation of the 5% quantile estimator. Motivate and explain your choice of bootstrap method.
- Now, repeat part b. but use as a bootstrap method a parametric bootstrap based on a lognormal distribution with suitably estimated parameters.<sup>1</sup> Compare the resulting estimate of the standard deviation of the 5%-quantile estimator with the one found in part b. Refer to the theory of Lecture 5, to explain what went wrong.
- Reprogram part c. so that everything, i.e. the calculation of the standard deviations, the realizations of (e.g.  $B = 1000$ ) bootstrap samples and the respective quantiles, 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 `sd`, `replicate`, `quantile`, `rlnorm`, `mean`.

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

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

- Determine a two-sided 95%-bootstrap confidence interval for the median of the underlying distribution of SDRP. Take  $B$  sufficiently large.
- Repeat part a with the mean instead of the median.
- Compare the answers of a. and b. Which estimator of location do you prefer and why?

---

<sup>1</sup>For a lognormal distribution, in R, parameters relate to the mean and the standard deviation of the logarithmized data. One may thus choose to use the sample mean and sample standard deviation of the logarithmized data as parameter estimates.

<sup>2</sup>For importing the data use the command `source("thromboglobulin.txt")`.

<sup>3</sup>see `help(attach)`

- d. Determine a 90%-bootstrap confidence interval for the difference in median between the two groups SDRP and CTRP. What can you conclude from this interval about the difference in median 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.

**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`<sup>4</sup>. 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:* you may use the following representation:

$$\tilde{D} = \sup_x \left| \hat{F}(x) - \Phi\left(\frac{x - \bar{X}}{S}\right) \right| = \max_i \max \left( \left| \frac{i-1}{n} - \Phi\left(\frac{X_{(i)} - \bar{X}}{S}\right) \right|, \left| \frac{i}{n} - \Phi\left(\frac{X_{(i)} - \bar{X}}{S}\right) \right| \right)$$

- 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  $\tilde{D}_n$  and the significance level  $\alpha = 5\%$ .

*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 for `dist` (distribution) and `par` (parameters). Perhaps the bootstrap version of the  $t$ -test statistic  $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ , i.e.  $t = \frac{\bar{X}^* - \bar{X}}{S^*/\sqrt{n}}$ , might give you an idea of what an appropriate bootstrap version of the Kolmogorov-Smirnov test statistic could look like. Here  $\bar{X}^*$  and  $S^*$  are respectively the sample mean and standard deviation of a bootstrapped sample. 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  $\tilde{D}_{n,i}^*$ ,  $i = 1, \dots, B$  (e.g.  $B = 1000$ ). Finally, use these bootstrapped values of the test statistic to find a suitable  $p$ -value of the test. Use the (default) argument `alternative = two.sided`; yet, the test is right-tailed, i.e. the null hypothesis is rejected for sufficiently large values of the test statistic.

- 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.

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