SDA 2019 — Assignment 3

For these exercises you can use the *R*-functions `ks.test` and `shapiro.test`, and the function `chisquare` that can be found on the Canvas page for this assignment. (The *R*-function `chisq.test` should *not* be used for chi-square tests for goodness of fit.) Investigate these functions before using them. When performing a statistical test, state the null and alternative hypothesis, present the test statistic and its distribution under the null hypothesis, give the value of the test statistic, the critical region or the p-value and the chosen significance level, and formulate the conclusion of the test.

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 3.1 The file `body.dat.txt` contains several body measurements (and additional information) of 507 individuals (mainly) in their twenties and thirties, all of them doing sport exercises for several hours per week. In this exercise, we focus on the calf and ankle girths (in cm; columns 19 and 20, respectively; averages of left and right leg girths were taken). For exercise parts a.-d., use only the first 70 rows of the dataset.

a. Investigate whether or not the calf and ankle measurements are from the same location-scale family. Use the function `qqplot` for a two sample QQ-plot.

b. Find for both vectors an appropriate distribution, each with its own parameters.

c. Investigate the normality of the differences between calf and ankle measurements (without using a hypothesis test). Does your answer on the question whether the differences are normally distributed or not, follow from your answer in part b already?

d. Test the normality of the differences using the Shapiro–Wilk test.

e. For this last part, include all *males'* ankle measurements (i.e. all 247 first rows). Create a Q-Q plot of the measurements and conduct another Shapiro-Wilk test on all these values. What is your interpretation of both outcomes? Do they contradict each other?

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

Exercise 3.2 on the next page!

Exercise 3.2

a. Explore the sample in `sample2019.txt` graphically and find an appropriate distribution from which this sample could have been drawn. Indicate location and scale as well.

b. Test (at level $\alpha = 5\%$) whether the sample originates from the Gamma distribution[1] with scale parameter $\theta = 2.2$ and shape parameter $k = 2$ using the Kolmogorov–Smirnov test.

c. Do the same as in part b, but now use the chi-square test for testing the goodness of fit. Choose the arguments of the function `chisquare` so that the condition for the rule of thumb (see syllabus) is fulfilled.
Hint: the function `qgamma` *could be useful for ensuring the rule of thumb condition.*

d. Do you find the results of parts b and c to agree with your results from part a?

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

[1] A gamma-distributed random variable $X \sim \Gamma(k, \theta)$ with scale $\theta > 0$ and shape $k > 0$ has the density function $x \mapsto \frac{x^{k-1}\exp(-x/\theta)}{\Gamma(k)\theta^k}$, where Γ denotes the gamma function.