SDA 2021 — Assignment 2

Note: The Exercises 2.2 and parts of Exercise 2.3 can only be solved after the lecture on Feb. 17 or after you have read the remainder of Chapter 3 in the syllabus!

For these exercises you can use the *R*-functions `summary`, `range` and `IQR`. For QQ-plots you may use `qqnorm`, as well as the local functions `qqt`, `qqlnorm`, `qqchisq`, `qqlogis`, `qqexp`, `qqunif` and `qqcauchy` which can be found on the Canvas page (`functions_Ch3.txt`)[1].

Also, 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.

Note: to indicate a normal distribution with expectation 2 and *variance* 25, we use the notation $\mathcal{N}(2,25)$, whereas *R* uses the parameters `mean=2`, and `sd=5` for this normal distribution.

When performing a statistical test, state the null and alternative hypothesis, present the test statistic and its distribution under the null hypothesis (if it is a well-known distribution), 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 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**.

General information on loading executable R code from a .txt file:
You can load the data that should be analyzed in Exercises 2.1 and 2.2 by performing the following steps:

i) save the file `sample2021.txt` to some directory called `thedirectory` with path `path`[2],

ii) set the working directory to `thedirectory` by using the command `setwd("path")` (instead of `path`, you obviously need to fill in the correct path on your computer!),

iii) finally run the code in the file `sample2021.txt` using the command `source("sample2021.txt")`.

The `i`th component of a list called `listname` can be extracted using `listname[[i]]`.

General information on loading data from a .txt file into your workspace:
Use the command `data <- read.table("path/....txt")`[2].

[1] These functions can be loaded in exactly the same way as the code from the file `sample2021.txt`, see the General information on page 1

[2] For Windows the path is usually `C:/.../thedirectory`, for Mac the path is usually `/Users/.../thedirectory`.

Exercise 2.1

a. Make plots of the quantile functions (that is, make 'true' QQ-plots as in Figure 3.4 of the syllabus or Slide "Quantiles of F and $F_{a,b}$ (2)" of Lecture 2) of the following pairs of distributions:

   I. uniform(1,2) – standard normal.

   II. normal with mean 1 and variance 4 – t_3.

   III. t_{10} – log normal with meanlog = 0.1, sdlog = 0.4.

   Comment on each plot on the heaviness of the tails of the two distributions. The tails of a distribution can be seen as the relative height of its density $f(x)$ for $x \to \pm\infty$. For example, the right tail of an exponential distribution is heavier than the right tail of a uniform distribution (which vanish for finite x). You could create for your own use some density plots to get a better idea of the tail behaviour of the different distributions (none of these density plots should be handed in).

   Note: for this exercise you should not generate random samples. Instead, use the true quantile functions for both the x-axis and the y-axis (for example, the R-function `qnorm` *can be used for computation of the quantiles of a normal distribution). For plotting the function* `plot` *should be used, not the function* `qqplot`.

The functions `qqnorm`, `qqt`, `qqlnorm`, etc., can be used to make QQ-plots for the location-scale families of the normal, t, lognormal distributions, etc., respectively. The argument `df` in `qqt` and `qqchisq` is used to set the number of degrees of freedom in the t-distribution and the χ^2-distribution.

b. To get an idea of what QQ-plots look like, test one or more of these functions by generating one or more samples from a distribution and making QQ-plots on your screen. None of these plots need to be handed in.

c. Investigate the data `sample2021a` in `sample2021.txt` with the given functions for making QQ-plots and find an appropriate distribution for this data set. Apart from giving a proper location-scale family (e.g., "normally distributed"), also give values for the location and scale parameters. (e.g., "normally distributed with location 2 and scale 5", or "$\mathcal{N}(2,25)$ distributed").

   Hints: 1. Using the commands `qqline` *or* `abline` *can be helpful! (See Lecture 2 for more details.) Note that slope and intercept of* `qqline` *are not the parameters a and b of the location-scale family.*
   2. The location and scale parameters in your location-scale family $F_{a,b}$ are $a \in \mathbb{R}$ and $b > 0$, respectively, and not necessarily equal to the parameter(s) of the underlying parametric distribution family. For instance, the exponential distribution is usually only parametrized by a so-called rate-parameter, where rate=1/scale; only the location-scale family also introduces a new, location-parameter.

Hand in: your plots and comments for part a and plots of relevant graphs of part c, as well as a motivation for your trials and your final conclusion for part c.

Exercise 2.2

a. Explore the sample `sample2021b` in `sample2021.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[3] with scale parameter $\theta = 2.1$ and shape parameter $k = 1.9$. For this, use 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. Explain whether the results from parts b and c agree. If you find they do not agree, find a reason why this might be so.

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

Exercise 2.3 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 this exercise we only use the first 247 rows of the dataset which correspond to all male individuals.

a. Make histograms and boxplots of the calf and ankle measurements and conclude whether both data distributions have approximately the same shape.

b. 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.

c. Without the use of hypothesis tests, find for each of the two datasets of calf and ankle measurements an appropriate distribution, each with its own parameters.

d. Investigate the normality of the differences between calf and ankle measurements (without using a hypothesis test).

e. Use the Shapiro–Wilk test to test the normality of the differences.

f. Compare the outcomes of the goodness-of-fit tests for normality for the full sample of calf measurements and for the first 50 calf measurements, and also use histograms of these two samples to complement your analyses. Find a possible explanation for these outcomes of the tests.

Note: in general for testing normality based on the Shapiro-Wilk test, one would just use the full sample and conduct the test only once (per sample). This exercise is just meant to gives us a better understanding of how goodness-of-fit tests work. In this exercise we assume that the first 50 entries in the calf measurements sample are representative for the full sample.

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

[3] 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.
Note: sometimes a different parametrization is used with parameters *rate* $= 1/\theta$ and *shape* k. Find out what is required here and make certain you use the correct parametrization!