SDA 2022 — Assignment 2

Note: Parts of Exercises 2.3 and 2.4 can only be solved after the lecture on Feb. 23 or after you have read the remainder of Chapter 3 in the syllabus!

For the exercises of Assignment 2 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`, `qqcauchy`, `qgompertz`, `dgompertz`, `pgompertz`, `rgompertz`, 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: How to load executable R code from a .txt file:
You can load the data that should be analyzed in Exercises 2.2 and 2.3 by performing the following steps:

i) save the file `sample2022.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 `sample2022.txt` using the command `source("sample2022.txt")`.

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

General information: How to load data into your workspace:
Use one of the following `read` commands, depending on your data type:
`data <- read.table("path/....txt")`[2],
`data <- read.csv("path/....csv")`[2],
`install.packages("xlsx")`
`library(xlsx)`
`data <- read.xlsx("path/....xlsx", sheetIndex=1)`[2].

[1] These functions can be loaded in exactly the same way as the code from the file `sample2022.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`.

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.

Exercise 2.1 (Class Exercise on Feb. 18) To get an idea of what QQ-plots look like and how to use the commands for making QQ-plots, you will apply some of these functions to some samples drawn from different distributions.

Exercise 2.2

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. standard normal against exponential with rate 3.

   II. normal with mean -1 and variance 9 against t_4.

   III. t_6 against chi-squared with 3 degrees of freedom.

   Comment for 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 some large $x > 0$). 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`.

b. Investigate the data `sample2022a` in `sample2022.txt` with the given functions for making QQ-plots and find an appropriate distribution for this data set. Apart from specifying a suitable location-scale family (e.g., "normal distributions"), also give values for the location and scale parameters. (e.g., "normal distribution with location 2 and scale 5", or "$\mathcal{N}$(2,25) distributed", or "$2 + 5 \cdot Exp(1)$, i.e. $Exp(1/5)$-distributed, shifted by 2").

   Hint: Using the commands `qqline` *and* `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.*

Hand in: your plots and comments for part a and plots of relevant graphs of part b, as well as a motivation for your trials and your final conclusion for part b. (Not all trials need to be documented, just the ones that really led you to the final model.)

General information on the .RData file to be submitted to the dummy assignment "RData A2" on Canvas:
Create in R a list `mylist` that contains the required entries as specified in the exercise(s) which are marked as "(partially) .RData file hand-in". You should store your list in an .RData file by using the R command `save(mylist, file="[PATH]/myfile2_[GROUP_NO].RData")`, where `[PATH]` stands for the path on your computer where you wish to save the .RData file and `[GROUP_NO]` stands for the number of the assignment group you have chosen on Canvas). **In any case**, `mylist` must have the entry `stud_no`: a vector that contains the student numbers of you and your group partner.

Exercise 2.3 (partially .RData file hand-in)

a. Explore the sample `sample2022b` in `sample2022.txt` graphically and find an appropriate distribution (i.e. a specific member of a location-scale family) 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 Gompertz distribution[3] with scale parameter $b = 0.2$ and shape parameter $\eta = 1$. Use the Kolmogorov–Smirnov test for this.

c. Do the same as in part b, but now use a 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 `qgompertz` *could be useful for ensuring the rule of thumb condition.*

d. Explain whether the results from parts b and c agree, and interpret the results. If you find they do not agree, find a reason why this might be so.

Hand in:
a.: relevant graphs, results, and comments **in the main report**.
b.: the following entries of your list `mylist` in R, which will be stored in your .RData file:

- `KS-score`: the value of the Kolmogorov-Smirnov test statistic applied to the data,
- `KS-p-value`: the resulting p-value of the Kolmogorov-Smirnov test,
- `KS-reject = TRUE` or `FALSE`: the test decision, whether the null hypothesis is rejected.

c.: the following entries of your list `mylist` in R, which will be stored in your .RData file:

- `Chisq-breaks`: the breaks, i.e. a vector specifying the lower and upper boundaries of the intervals used by the Chisquared-goodness-of-fit test; see Lecture 3 for details.
- `Chisq-score`: the value of the Kolmogorov-Smirnov test statistic applied to the data,
- `Chisq-p-value`: the resulting p-value of the Kolmogorov-Smirnov test,
- `Chisq-reject = TRUE` or `FALSE`: the test decision, whether the null hypothesis is rejected.

d.: your answers, with reference to the test outcomes in b. and c. (repeat here what the outcomes were), **in the main report**.

[3] A Gompertz-distributed random variable $X \sim GO(\eta, b)$ with scale $b > 0$ and shape $\eta > 0$ has the density function $x \mapsto b\eta \exp(\eta + bx + e^{bx})$.

Exercise 2.4 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 ankle girths (in cm; column 20) and the body mass indices (BMIs), which are weight (in kg) divided by the squared height (in meter). Column 23 and 24 respectively contain the weight (in kg) and height **(in cm)** of the individuals. For this exercise we only use the first 247 rows of the dataset which correspond to all male individuals.

a. First obtain the BMIs from the weights and heights. Next, make histograms and boxplots of the BMIs and ankle measurements and conclude whether both data distributions have approximately the same shape.

b. Investigate whether or not the BMIs and ankle measurements could be 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 BMIs and ankle measurements appropriate distributions, as members of certain location-scale families.

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

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

f. Compare the outcomes of the goodness-of-fit tests for normality for the full sample of BMIs and for the first 50 BMIs, 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 give us a better understanding of how goodness-of-fit tests work. In this exercise we assume that the first 50 BMIs sample are representative for all others.

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