SDA Group Submission Assignment 2

CD

For your group submission, upload one .pdf in Canvas. Please refer to the R Markdown cheat sheet and the R Markdown Submission template (to be found in Canvas) to get started with R Markdown 😀

If you follow all installation instructions given in Canvas and still have big issues with R Markdown installation, you can exceptionally submit a .pdf generated in a different way (e.g., by copying code 'manually' into a text editor), which should still include all relevant explanations, code snippets and graphics.

Always remember that there are several ways to code something, and that the useful `R` functions provided for each task are just some suggestions (that is, you can use your favorites even if they are not listed). Finally, the symbol ϟ indicates something generally worthy of attention. Have fun!

Exercise 1

a. Consider the following pairs of distributions:

   i. t_3 (that is, a t distribution with 3 degrees of freedom) and $\mathcal{N}(0, 2)$ (that is, a normal distribution with mean 0 and *variance* 2)
   ii. χ^2_{25} (that is, a chi-squared distribution with 25 degrees of freedom) and χ^2_4;
   iii. $\Gamma(7, 3/4)$ (that is, a Gamma distribution with shape 7 and rate $3/4$) and $\text{Exp}(3/4)$ (that is, an exponential distribution with rate $3/4$).

   Create a 2×3 grid of plots using `par(mfrow = c(2, 3))` (it will fill *row-wise*), consisting of:

   - for each pair of distributions given above, a *true* QQ-plots as in Figure 3.4 of the Syllabus;
   - below each *true* QQ-plot, a 'density-comparison' plot with the corresponding pair of theoretical densities, each plotted as a line (use two different colors). Remember to set the y-axis interval such that nothing is cropped vertically, and the x-axis interval in a suitable way. You can make use of a legend (see `?"legend"`) to clarify which line corresponds to which theoretical density, or you can make good use of axes' labels and/or of a main title.

   Hint: here, you should not generate random samples. Instead, use the *true* quantile function of each distribution (such as `qnorm` for the normal distribution) and apply it on a sequence of so-called *probability points* (that is, points within $(0, 1)$ – boundaries excluded). ϟ Then, plot the quantiles of a distribution against those of the other using `plot` (and not `qqplot`).

b. Investigate the data in `mysterious_sample.RDS`. First, read the file into `R` (see chunk below). Then, create QQ-plots by comparing sample quantiles to the theoretical quantiles of the uniform distribution, the extreme value distribution and the logistic distribution.

   This can be easily done in `R` by installing the package `EnvStats`[1] and then using its function `qqPlot`.

   In the following chunk, you find an example on how to use the function to compare the sample quantiles of a sample `mystery` to the theoretical quantiles of $\mathcal{U}(0, 1)$, $\text{EVD}(0, 1)$ and $\text{Logistic}(0, 1)$. You may still need/want to set appropriate plotting options of your choice.

[1] Do **not** include the installation command in your .Rmd: this would install the package every time you knit 😬 Rather, run the command `install.packages("EnvStats")` in console only. Alternatively, install the package through RStudio: click on **Packages** in the menu of the bottom-right pane (this pane menu is typically set on **Plots**), then click on **Install** and start typing the package name in the text field, select the desired package from the drop-down menu and finally click on **Install**. Recall that you only need to install a package once, afterwards to use it you just need to load it with the command `library(...)`.

You can use the given theoretical distributions to answer the question: which distribution represents an appropriate fit for the given data? Apart from naming the distribution, also give estimates for the values of a and b (the intercept and slope of the straight line in the chosen QQ-plot). To that, as explained in the slides for Lecture 2, one should make use of the mean and standard deviation of the sample, as well as of their theoretical counterparts for the chosen distribution (expectation and *variance* for the three proposed distributions are reported below).

Distribution	Parameters	Expectation	Variance
$\mathcal{U}(a, b)$	$-\infty < a < b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\text{EVD}(\eta, \theta)$	$\eta \in \mathbb{R}$, $\theta > 0$	$\eta + 0.5772 \cdot \theta$	$\frac{\pi^2}{6}\theta^2$
$\text{Logistic}(\mu, \sigma)$	$\mu \in \mathbb{R}$, $\sigma > 0$	μ	$\frac{\pi^2}{3}\sigma^2$

Hint: ϟ The `.RDS` file should be saved in the same folder as your .Rmd, and that folder should be set as working directory: one way to do this is from the RStudio menu, where you can select **Session** and then **Set Working Directory**, from where you can choose the folder where the `.RDS` and .Rmd are. Once you find estimates for a and b, you can use `abline(a = ..., b = ...)` to add a a line to the appropriate QQ-plot to confirm/disprove whether your estimates are suitable (no need to hand this extra plot(s) in).

```
library(EnvStats)
# Read-in data
mystery <- readRDS("mysterious_sample.RDS")

# Compare x with U(0, 1)
qqPlot(mystery, distribution = "unif",
       param.list = list(min = 0, max = 1), add.line = TRUE)

# Compare x with Cauchy(0, 1)
qqPlot(mystery, distribution = "evd",
       param.list = list(location = 0, scale = 1), add.line = TRUE)

# Compare x with Logistic(0, 1)
qqPlot(mystery, distribution = "logis",
       param.list = list(location = 0, scale = 1), add.line = TRUE)
```

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The graphical output required by a. and b.
- The answer to the final question in b.

Useful `R` functions

`ppoints`, `par(mfrow = c(i, j))`, `plot`, `curve`, `qnorm`, `qt`, `qchisq`, `qgamma`, `qexp`, `dnorm`, `dt`, `dchisq`, `dgamma`, `dexp`, `readRDS`, `EnvStats::qqPlot`, `mean`, `sd`...

Exercise 2

a. As you may know, there are parakeets in Amsterdam. They have been establishing themselves in the city since the late '70s, it is estimated. In this exercise, you will work with *fictitious* (unfortunately) censuses of these parakeets. The data set in file `parakeets.RDS` (Exercise 1 gives indication on how to read `.RDS` files into `R`) contains the following three columns:

- `Name`: parakeet name;
- `Year`: year in which the parakeet was observed and weighed;
- `Weight`: parakeet weight (in grams).

Plot two histograms (scaled to density) of the parakeet weight distribution side-by-side: one considering *only* the parakeets weighed in 1980 and one *only* those weighed 2015. Set the x-axis interval in each histogram to $(100, 140)$ for an easier comparison of the two sample distributions. What do you observe?

b. Use the function `qqplot` to create a two-sample QQ-plot, where one sample consists of the 1980 weights and one of the 2015 weights. Could the two sample originate from the same location-scale family?

c. Investigate the normality of the two samples graphically with `qqnorm` and `qqline` (plot the two QQ-plots side-by-side). Do the samples seem to originate from a normal distribution?

d. Investigate the normality of the two samples by using the Shapiro-Wilk test with significance level $\alpha = 5\%$. Comment of the result of each test: is the null hypothesis rejected? Explain why/why not. How can you interpret these results in terms of normality of each sample?

e. How do the test results in d. compare with what you observed graphically? What could be a possible explanation of these test outcomes?

Hint: although in this exercise it is not necessary for you to show the `R` code which generates the plots, always remember to make your plots are readable as possible by setting main titles, labels etc. in an appropriate way wherever necessary.

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The graphical output required by a., b. and c.
- The `R` function calls and corresponding outputs for the tests in d.
- The comments required in each point
- Your explanation for the last question in e.

Useful `R` functions

`readRDS`, `par(mfrow = c(i, j))`, `hist`, `qqplot`, `qqnorm`, `qqline`, `shapiro.test`,...