Practice exam

Statistical Computing with R (2023-24)

Your data (fill this section in!)

- Name:
- Surname:
- Student number:

Important: please **hand in this document to the invigilators at the end of the exam** (even if you decide to leave earlier)!

Grading table (leave this section empty)

Exercise 1		Report quality and clarity of answers	
Exercise 2			
Exercise 3		Final score	

Instructions

Please **read these instructions carefully** before starting to work on the exam.

How to organize your solutions

1. Use R Markdown to write your solutions to the exam. Include your name and student number in the "Author" field, choose an appropriate title for the document, and name your `.Rmd` file as follows: `surname_name_practiceExam.Rmd`, where `surname` is your surname and `name` your name.
2. Compile the document to `pdf`. At the end of the exam, **submit both the `Rmd` file and the compiled `pdf`** with your solutions through Brightspace
3. Use section and subsection headers to structure the document, so that it is clear which exercise (section header) and question (subsection header) you are answering.
4. Put your `R` code in code chunks. Keep the argument `echo = TRUE` (= don't hide your code in the compiled pdf!).
5. Write your answers inline (= outside code chunks); please don't include your answers as comments inside code chunks.
6. When relevant, please **clearly justify your answers**. The clarity of your answers and report quality will contribute to the exam grade!

What can you use / not use during the exam?

- **You can use the lecture slides** (without notes / annotations) during the exam. You are **not allowed to use the material from the coding sessions / practicals**. You are also not allowed to use any other document (books, notes, own scripts / files, ...).
- You can use `R` and RStudio to consult help pages of relevant functions (e.g., `?mean`, `?substr`)
- You are not allowed to ask for help from your colleagues. If you run into technical problems, please contact the instructor / TAs present in the room instead.
- You are not allowed to use any form of generative AI (ChatGPT, Google Bard, ...) to solve the exam.
- You can use your internet connection to submit your solutions to the exam, and (if needed) to install `R` packages and download data files that are necessary to solve the exam. Any other use of your internet connection (consulting internet to find solutions, communicating with others, etc.) is not permitted.
- If discovered, suspected cases of cheating will be reported to the Board of Examiners. So: be honest!

A final request

This exam has been designed to closely match the course contents, and test your knowledge of such contents. For this reason, you are expected to **solve the exercises using as much as possible the functions and packages discussed during the course**, rather than alternative packages / approaches that were not covered in this course (if, nevertheless, you decide to use alternative packages / approaches, please include an explanation of why you are doing it).

Good luck!

Exercise 1

The file `cartype.csv` dataset contains data on car registrations by engine type from 2000 onwards for a selected list of countries. It considers 4 different types of engines: petroleum; diesel; hybrid; and fully electric battery vehicles.

1. Download the file from Brightspace, and import it in `R`. Is the dataset in wide or long format?
2. Select the data from the following 5 countries: France, Germany, Italy, Spain and United Kingdom. For the rest of the exercise, you can ignore the data from all other countries.
3. Compute 4 variables that contain the proportion of registered vehicles for each engine type in a given year.
4. In which country were diesel cars more popular in 2008? And in 2018?
5. Create a pie chart showing the distribution of car registrations by engine type in Germany in 2005. Do the same for 2018. How did things change from 2005 to 2018?
6. Create a spaghetti (trajectory) plot that displays the change over time of the percentage of hybrid vehicles across the 5 countries. Which country/countries has/have seen a faster diffusion of hybrid vehicles?
7. Create a spaghetti (trajectory) plot that displays the change over time of registrations by type of engine in France. What are the main temporal patterns that you can observe?

Exercise 2

A random variable X follows the following density function with parameters $k \geq 2$ and $\lambda \geq 0$:

$$f(x; k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, x \geq 0.$$

1. Draw the density $f(x; k, \lambda)$ over the interval $[0, 10]$ when $k = 7$ and $\lambda = 2$.
2. Let $k = 3$ and $\lambda = 4$. Use a numeric optimization method to find the mode of $f(x)$ for the given k and λ. Tip: you may restrict your search for the mode to the interval $x \in [0.01, 30]$.
3. Let $k = 2.3$ and $\lambda = 5.7$. Use a numeric optimization method to find the mode of $f(x)$ for the given k and λ. Tip: you may restrict your search for the mode to the interval $x \in [0.01, 30]$.
4. If possible, maximize $f(x; k, \lambda)$ analytically to obtain a closed-form expression for the mode. You may answer this question either using pen and paper, or directly in `R` Markdown (typing your formulas with LaTeX).
5. Using the expression that you derived at (4), write a function that computes the mode of X as a function of k and λ. Use such function to double-check the results that you obtained at points (2) and (3). Do you obtain the same results? Can you mention at least one advantage, and at least one disadvantage, of solving the problem numerically instead of analytically?

Exercise 3

1. Create a function called `getSummaries` that given a numeric input vector x returns a data frame with the following information: sample size, mean, median and variance. The output should look like this:

Statistic	Value
Sample size	24
Mean	-3.5443
Median	-3.0003
Variance	2.0112

2. Add to `getSummaries` a `digits` argument that determines the number of digits that the mean, median and variance should be rounded to.
3. Add to `getSummaries` a further argument called `ignoreNAs` that determines whether to ignore `NA`s:

a. if the argument is `FALSE`, the function should produce a warning message, and only return the sample size;
b. if the argument is `TRUE`, the function should return: the original sample size, the sample size after removal of the `NA`s, mean, median and variance.

4. Continue editing `getSummaries` so that if the user supplies a matrix instead of a vector, the function computes the summary statistics mentioned at point (1) separately for each column of the matrix.
5. Edit `getSummaries` so that it returns an error if the input `x` is neither a numeric vector nor a matrix.

Now let

```
set.seed(3078)
x1 = rhyper(300, 5, 20, 10)
x2 = c(rep(cars$speed, 3), rep(NA, 50), rep(cars$dist, 2))
x3 = cbind(x1, x2)
```

6. Apply `getSummaries` to `x1`, `x2` and `x3` setting the following arguments: `digits = 2` and `ignoreNAs = T`.
7. Apply `getSummaries` to `x1`, `x2` and `x3` setting the following arguments: `digits = 4` and `ignoreNAs = F`.