# Assignment 3

Statistical Computing with R, 2023-24

## Introduction

This assignment consists of 2 exercises that have been designed to test the skills you have learned in the first 11 lectures of Statistical Computing with `R`.

When solving the assignment, please bear in mind the following points:

1. you are warmly encouraged to use the packages and functions that you have encountered in the course so far, as well as the ones given as hints in this assignment;
2. the assignment is optional and not graded. You are allowed to discuss your solutions and collaborate with classmates, but please prepare your own solutions, submitting code and answers that you wrote by yourself - this way we can give you feedback about your thought process and code style, which will help you assess your level of preparation so far and will be useful to improve your skills before the exam;
3. try to structure your solutions clearly, and write tidy and commented code.

### Preparing your solutions

Please use `R` Markdown to prepare and hand in your solutions:

1. use `R` Markdown to write your solutions to the assignment. Name your `.Rmd` file as follows: `surname_name_SCwR_A3.Rmd`, where `surname` is your surname, and `name` your name;
2. use section and subsection headers to structure the document, so that it is clear which exercise and question you are answering;
3. put your `R` code in code chunks. Keep the argument `echo = TRUE` (= don't hide your code in the compiled pdf!);
4. write your answers inline (= outside code chunks); please don't include your answers as comments inside code chunks!
5. compile your solutions as `.pdf`. If you have issues in compiling your document as pdf, you can reach out to the TAs for help!

### Submission

The **deadline** for submissions is **December 5 (Tuesday) at 23:59**. To hand in your solutions, submit both the `Rmd` source file and the compiled `pdf` report through Brightspace.

### Doubts / questions?

If you have any questions about the assignment, feel free to ask the TAs during the coding sessions, the question hour or via email to statcompr[at]gmail.com.

Good luck, and have fun!

# Exercise 1

The `penguins` dataset from the `R` package `palmerpenguins` contains variables measured on different species of penguins in the Antarctic Peninsula.

1. Read the help page of the `penguins` dataset. What type of object `penguins` is? Load the object in `R`, and convert it to a data frame.
2. Compute the (joint) frequency distribution of penguins by species and island. How many penguin species are present in this data frame? Which penguin species are present on the different islands?
3. Assuming that the distribution of `bill_length_mm` is (approximately) normal for each species, test the null hypothesis that $H_0 : \mu_C \geq \mu_G$, where $\mu_G$ is the expected value of `bill_length_mm` for Gentoo penguins, and $\mu_C$ is the expected value of `bill_length_mm` for Chinstrap penguins. Show the result of the test, and set $\alpha = 0.05$. What do you conclude?

# Exercise 2

Consider again the `penguins` data frame that you already used in exercise 1.

1. Are there penguins for which `species` or `body_mass` are missing? If yes, remove such penguins from the dataset.
2. Using functions from either base `R` or from `ggplot2`[1], create a density plot that compares the distribution of weight for the different species. What can you observe?

## Estimation of a mixture model with 3 components

Now assume that the distribution of weight $W$ in kilograms (convert `body_mass_g` from grams to kilograms, and use the newly created variable for the rest of the exercise!) follows a mixture of 3 normal distributions, whose components are denoted by

$$W_1 \sim N(\mu_1, \sigma_1), W_2 \sim N(\mu_2, \sigma_2) \text{ and } W_3 \sim N(\mu_3, \sigma_3)$$

with unknown weights $\pi_1, \pi_2, \pi_3$ and unknown means and standard deviations $\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3$.

3. Write a function that evaluates the negative log-likelihood of the mixture model.
4. Set your random seed with `set.seed(stn)`, where `stn` is your student number. Implement an EM algorithm that estimates the mixture model, and apply it to $W$ (body mass in kg), taking the following instructions into account:

- Run the algorithm from 3 (different) random starting points, i.e. using different $\pi_j^{(1)}$ and $p_{ij}^{(1)}$ for each algorithm run. **Explain how you have chosen such starting points**.
- Let each run of the EM algorithm run for **500 iterations**.
- Please include all relevant code in the `.Rmd` file with your solutions (keeping `echo = T` in all code chunks), rather than sourcing code from external `R` scripts.

## Selecting an appropriate ML solution

5. Compare the output of the 3 runs of the EM algorithm. Which of the 3 solutions should you pick? **Why?** For the rest of the exercise, make use of such solution.
6. Check if after 500 iterations the algorithm appears to have converged.
7. What are your maximum likelihood estimates of $\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3$? Are the weights of the 3 components similar? And are your estimates of $\mu_1, \mu_2, \mu_3$ “close” to, or very different from, the mean weight for each penguin species?

[1] we will introduce `ggplot2` in Lecture 12

## Classification

8. Estimate the probability that each observation (= penguin) comes from $W_1$, $W_2$ and $W_3$. Draw 3 histograms (with `breaks = 20`, or anyway a sufficient number of bins) that display the distributions of $W_1$, $W_2$ and $W_3$.
9. Create a variable called `predictedComponent` that takes values 1, 2 or 3, and that represents your best guess of whether each individual comes from $W_1$, $W_2$ or $W_3$.

Let's now try to relate `predictedComponent` to the variables `sex` and `species`.

10. Compute the (joint) frequency distribution of the variables `sex` and `predictedComponent`, as well as of `species` and `predictedComponent`.
11. If a penguin has `predictedComponent = 1`, would you guess that it is a male or a female? **Why?** What about `predictedComponent = 2`, and `predictedComponent = 3`? How accurate would your guesses be?
12. What about species? Which species would you guess a penguin with `predictedComponent = 1` is from? **Why?** What about `predictedComponent = 2`, and `predictedComponent = 3`?