Unsupervised Learning and Evolutionary Computation Using R

Winter Term 2024/2025

Exercise Sheet 3 (November, 11, 2024)

Exercise 1 (Recap: normal distribution)
Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be independent and identically distributed random variables. Show that for

$$Y := \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

it holds that $E(Y) = 0$ and $\text{Var}(Y) = 1$.

Exercise 2 (QQ-Plots)
Consider the *penguins* dataset from the package *palmerpenguins* which provides various measurements for a group of adult penguins in Antarctica. Below you are given 10 observations of this data set from which `NA` values have been removed (you can use the function `complete.cases()` for subsetting). Your task is to check those data of the variable `flipper_length_mm` for normality.

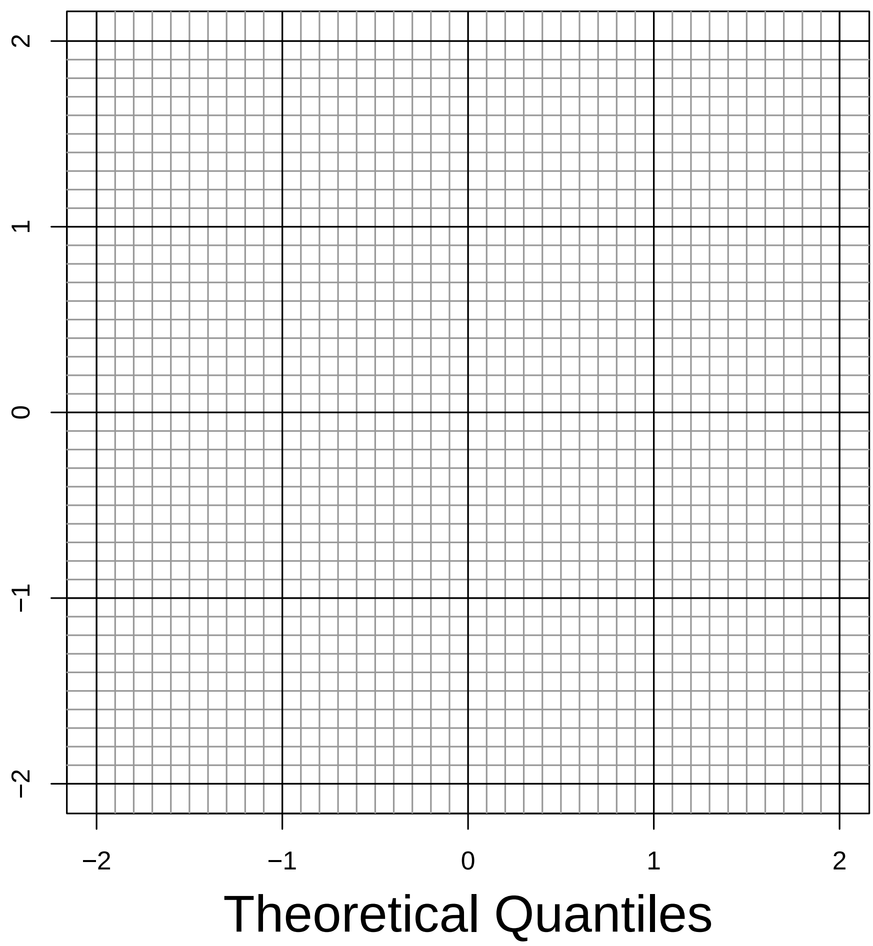
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm
1	Adelie	Torgersen	39.1	18.7	181
2	Adelie	Torgersen	39.5	17.4	186
3	Adelie	Torgersen	40.3	18	195
4	Adelie	Torgersen	36.7	19.3	193
5	Adelie	Torgersen	39.3	20.6	190
6	Adelie	Torgersen	38.9	17.8	181
7	Adelie	Torgersen	39.2	19.6	195
8	Adelie	Torgersen	41.1	17.6	182
9	Adelie	Torgersen	38.6	21.2	191
10	Adelie	Torgersen	34.6	21.1	198

a) Normalise the variable appropriately so that you can check for standard normal distribution (you can use R for this purpose). Provide the values of this variable `flipper_length_mm_norm`.

b) Fill the following table as the basis for generating the required data for the QQ-plot below. For identical values, randomly assign them to the related adjacent ranks (e.g., two identical values can have 3rd and 4th rank). Use R to find the required values for q (normal):

$flip_n$	ranks	j^*	q (normal)
	1		
	2		
	3		
	4		
	5		
	6		
	7		
	8		
	9		
	10		

c) Complete the QQ-plot below and insert the qq-line. Are you deciding for or against a possible normal distribution? Explain your reasoning.



Exercise 3 (Shapiro-Wilk Test Outlier Sensitivity)
Reproduce the box-plots from the lecture slides on the sensitivity of the Shapiro-Wilk normality test to a single outlier. To this end for each sample size $n \in \{100, 1\,000, 2\,500\}$ and each outlier $o \in \{4, 4.2, 4.4, \ldots, 5.8, 6\}$ repeat the following experiment 30 times:

- Sample n random values from an $\mathcal{N}(0, 1)$-distribution.
- Add the outlier o to the sample.
- Apply the Shapiro-Wilk test and store the p-value.

Plot the distribution of the p-values for each outlier split by the sample size n. Interpret the results.

Exercise 4 (χ^2-distribution properties)

1. Let $X_1, \ldots, X_p$ be independent identically $\mathcal{N}(\mu, \sigma^2)$-distributed random variables. Show that
$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2(p)$$
where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. Now let $U_i \sim \chi^2_{p_i}, i = 1, \ldots, l$ be l independent random variables. Show that
$$\sum_{i=1}^{l} U_i \sim \chi^2_{p_1+\ldots+p_l}.$$

Exercise 5 (Outlier Detection Study)
Load the `heptathlon` data set from the `HSAUR3` R package. Familiarise yourself with the data set, look for possible outliers in the data and interpret your findings

Exercise 6 ((Bi-variate) Normal Distribution ★)
Let the density of a bi-variate variable $Z = (X_1,\ X_2)^T$ be given by the following expression:
$$f_Z(x_1, x_2) = \frac{1}{4\pi \cdot \sqrt{1-\rho^2}} \cdot \left(\exp\left(-\frac{x_1^2 - 2x_1x_2\rho + x_2^2}{2 \cdot (1-\rho^2)}\right) + \exp\left(-\frac{x_1^2 + 2x_1x_2\rho + x_2^2}{2 \cdot (1-\rho^2)}\right)\right)$$
Proof that the marginal distributions of $f_Z(x_1, x_2)$ are $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ with
$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x_i^2}{2}\right), \quad i = 1, 2.$$

Hints:

- The marginal density is defined as follows: $f_{X_2}(x_2) = \int_{-\infty}^{+\infty} f_{(X_1,X_2)}(x_1, x_2)\, dx_1$ (analogous for the marginal density $f_{X_1}(x_1)$)
- Split the bivariate density into two terms and integrate each term on its own (or better: get rid of the two terms within the brackets by simplifying the density)
- Split the exponential terms into a product of two terms by making use of artificially adding $+\,(x_2\rho)^2 - (x_2\rho)^2$ in the numerator

- Also, try to use the properties of a normal distribution and densities in general!
- Don't be frustrated if you fail, this is not an easy task, but try to do your best!