Probability and Statistics
Lab assignment 4: Hypothesis testing

General comments:

- This is a team assignment to be submitted before **23:59 of 3 December 2025**; complete solution will give you **4** points.
- You have to submit both the **Rmd** source file and the **html** output (links to GitHub repositories are NOT accepted!)
- **At the beginning of the notebook, provide a work breakdown structure estimating the efforts of each team member.**
- For each task, include
  - the corresponding **R** code,
  - the statistics obtained (like sample mean or anything else you use to complete the task),
  - your conclusions (e.g. whether to accept or reject the hypothesis) and explanations
- You are allowed to yourself create **your teams of three**. The `id number` of your team, which is referred to in tasks, is calculated as the sum of last digits in your students' ids of all team members. Observe that the answers do depend on this `id number`.
- Take into account that not complying with these instructions may result in point deduction regardless of whether or not your implementation is correct.

The data for problems 1–3 are generated as follows: set

$$a_k := \{k \ln(k^2 n + \pi)\}, \qquad k \geq 1,$$

where $\{x\} := x - [x]$ is the fractional part of a number x and n is your `id number`. Sample realizations $X_1, \dots, X_{100}$ and $Y_1, \dots, Y_{50}$ from the hypothetical normal distributions $\mathscr{N}(\mu_1, \sigma_1^2)$ and $\mathscr{N}(\mu_2, \sigma_2^2)$ respectively are obtained as

$$\begin{aligned} x_k &= \Phi^{-1}(a_k), \qquad && k = 1, \dots, 100, \\ y_l &= \Phi^{-1}(a_{l+100}), \qquad && l = 1, \dots, 50, \end{aligned}$$

where Φ is the cumulative distribution function of $\mathscr{N}(0,1)$ and Φ^{-1} is its inverse.

In **R**, you can define a function f calculating a_k from k, then apply f to the whole list of k's to get the list `a.data` of a_k, and, finally get x_k and y_k by running `qnorm` on `a.data`.

Instructions: In problems 1–3, test H_0 vs H_1. To this end,

- point out what standard test you use and why;
- indicate the general form of the rejection region of the test H_0 vs H_1 of level 0.05;
- find out if H_0 should be rejected on the significance level 0.05;
- indicate the p-value of the test and comment whether you would reject H_0 for that value of p and why

Problem 1. $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$; $\quad \sigma_1^2 = \sigma_2^2 = 1$.

Problem 2. $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 > \sigma_2^2$; $\quad \mu_1$ and μ_2 are unknown.
Hint: this is the f-test; read the details in Ross, p. 321–323

Problem 3. Using Kolmogorov–Smirnov test in **R**, check if

(a) $\{x_k\}_{k=1}^{100}$ are normally distributed (with parameters calculated from the sample);

(b) $\{|x_k|\}_{k=1}^{100}$ are exponentially distributed with $\lambda = 1$;

(c) $\{x_k\}_{k=1}^{100}$ and $\{y_l\}_{l=1}^{50}$ have the same distributions.

Explain the main idea behind the KS test and comment on the outcomes of the test.

Note: $\{x_k\}_{k=1}^{100}$ means a set of a hundred x_k with corresponding indices here, not their fractional part.

Problem 4. In this task you'll practice fitting the regression line to some real-life features and analyzing the results. The file data.csv contains data on students, specifically their study time and corresponding marks. Your tasks are as follows:

(a) Create a scatter plot of Marks vs. Study Time and provide brief comments;

(b) Fit a linear regression model using marks as the dependent variable and study time as the independent variable. Explain shortly the process of deriving the regression equation;

(c) Evaluate the goodness-of-fit for the fitted line;

(d) Suggest a way to test whether the study time is significant in predicting marks. Find the corresponding test statistics, specify its distribution. Find the p-value of the test and make a conclusion;

(e) If Alice studies for approximately 8 hours, what grade can we predict for her?

(f) Suggest up to three ideas to potentially improve prediction accuracy;