Probability and Statistics
Lab assignment 4: Hypothesis testing

General comments:

- This is a team assignment to be submitted before **23:59 of 20 December 2022**; complete solution will give you **3** points.
- The assignment must be completed in **R** language for statistical computing (`https://www.r-project.org/`). It can be installed from the official site; RStudio (`https://www.rstudio.com/`) is a convenient GUI.
- You will need just a few basic **R** commands to complete the task. As a quick reference guide, use the official manual `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf` or help section of RStudio.
- The assignment must be prepared as an R `notebook` and submitted to `cms` along with converted **pdf** document.
- 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 whether to accept the null hypothesis.
- 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`.
- It is obligatory to include the list of your team members at the beginning of your **R** notebook and to explain who was responsible for which part.

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

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

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

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

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.