Supervised Machine Learning

Professors: Fernando Díaz - Sebastián Azócar

Assistants: Pablo Leiva, Barbara Contardo, Pablo Flores

Fall Semester 2025, Group Work 2

"By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."

Eliezer Yudkowsky

Instructions

1. You must carry out this work in groups of 3 or 4 people.
2. Cooperation between groups is **strictly prohibited**.
3. The deadline to submit the corresponding report and all backup files is until Thursday, October 30, at 23:59.
4. It is required that you submit your homework, ensuring that you include a comprehensive report along with a hyperlink to Google Colab.

Assignment: Monte Carlo Simulation – Comparing the Power and Size of Classical and Bootstrap t-Tests

In the lecture on *Resampling Methods: Bootstrapping* (see Lecture 8), we compared the performance of the classical Welch t-test with the bootstrap-based t-test for the difference in means between two samples drawn from non-normal populations. In this assignment, you will conduct a **Monte Carlo experiment** to evaluate and compare the empirical *size* and *power* of these two tests.
—

Context

Consider two independent populations, each with the same mean but different distributional shapes:

- **Population 1 (Left-skewed):** $X \sim \text{Exponential}(\lambda = 1)$
- **Population 2 (Heavy-tailed):** $Y \sim t(\text{df} = 3)$

After generation, both populations should be **mean-adjusted** so that they share the same empirical mean:

$$X' = X - \bar{X} + \bar{Z}, \qquad Y' = Y - \bar{Y} + \bar{Z},$$

where $\bar{Z}$ is the overall mean of the combined samples. This ensures that any differences found in your analysis are due to sampling variability and distributional shape rather than differences in true means.
From each adjusted population, draw small random samples:

$$n_1 = 30, \quad n_2 = 35.$$

You will compare two tests of the null hypothesis:

$$H_0 : \mu_X = \mu_Y \quad \text{against} \quad H_1 : \mu_X \neq \mu_Y.$$

—

(1) Classical Welch t-test

The Welch t-statistic is given by:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}},$$

where $\bar{X}$ and $\bar{Y}$ are the sample means and s_X^2, s_Y^2 are the sample variances.
The corresponding p-value is computed assuming that, under H_0, t follows a t-distribution with Welch–Satterthwaite degrees of freedom.
—

(2) Bootstrap t-test

For the bootstrap version, draw B resamples (e.g., $B = 1000$) with replacement from each sample, compute the mean difference for each bootstrap sample:

$$\Delta_b = \bar{X}_b^* - \bar{Y}_b^*,$$

and estimate the bootstrap standard error:

$$\mathrm{SE}_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\Delta_b - \bar{\Delta})^2},$$

where $\bar{\Delta}$ is the mean of the bootstrapped differences. The bootstrap t-statistic is then:

$$t_{boot} = \frac{\bar{X} - \bar{Y}}{\mathrm{SE}_{boot}}.$$

The empirical distribution of $\{\Delta_b\}$ can be used to compute the corresponding p-value.

—

(3) Monte Carlo Experiment

Perform a Monte Carlo simulation with $R = 1000$ replications under the following scenarios:

(a) **Size:** Generate samples from populations with the same mean (i.e., after mean adjustment, $\mu_X = \mu_Y$). Record the proportion of replications in which H_0 is rejected at $\alpha = 0.05$.

$$\text{Empirical Size} = \frac{\#(\text{Reject } H_0 \mid H_0 \text{ true})}{R}$$

(b) **Power:** Generate samples where $\mu_Y = \mu_X + \delta$ for a fixed $\delta > 0$ (e.g., $\delta = 0.5$). Record the proportion of rejections at $\alpha = 0.05$.

$$\text{Empirical Power} = \frac{\#(\text{Reject } H_0 \mid H_1 \text{ true})}{R}$$

—

(4) Tasks

- Implement both tests (Welch and bootstrap) in Python or R.
- Compute and report the empirical size and power for each test.
- Compare and discuss:
  - Which test better maintains the nominal size under non-normality?
  - Which test exhibits higher power when $\delta > 0$?
  - How do small sample sizes affect each method?
- Include summary tables and visualizations (histograms or boxplots) of the estimated power and size across simulations.

—

Deliverable

Submit a report (in PDF or Jupyter Notebook format) including:

1. A description of your simulation setup.
2. Code snippets used to generate results.
3. A table comparing empirical size and power for both tests.
4. A brief discussion interpreting your findings.
5. A link to Colab.