Statistics and Data Science 1 | Assignment 2

Hypothesis Testing & Statistical Inference with Python

Core topics:

Topic 5: Probability & Probability Distributions (background)

Topic 6: Statistical Inference

Topic 7: Hypothesis Testing

Topic 8: ANOVA & Regression as an Alternative to Tests (if applicable)

Weight: 40 points of 2nd attestation

Deadline: 02.02.2026

Format: Individual work

Objective

The goal of this assignment is to practice statistical inference and hypothesis testing using Python. Unlike Assignment 1, which focused on exploring and describing data, this assignment focuses on making statistical conclusions based on data. Students are expected to formulate hypotheses, apply appropriate statistical tests, and interpret the results correctly.

Concepts from Week 5 (probability and distributions) are used as theoretical background and do not require separate analysis. Topics from Week 6 and Week 7 form the core of this assignment.

Dataset choice (choose ONE)

Students may choose one dataset using one of the following options.

Option A — Reuse dataset from Assignment 1 You may reuse the same dataset analyzed in Assignment 1, provided that it allows hypothesis testing (for example, comparison between groups or analysis of relationships between numerical variables). This option is recommended, as it ensures continuity and deeper understanding of the data.	**Option B** — Use a new dataset You may choose a new dataset from open sources (e.g., Kaggle, open datasets), as long as: • the dataset contains numerical variables, • at least one clear statistical hypothesis can be formulated. In both cases, the dataset must be clearly described and appropriate for hypothesis testing.

Task 1 - Dataset description & research question (10 points)

1. Briefly describe the dataset:

- what the data represents,
- number of observations,
- variables used in the analysis.

2. Formulate at least ONE clear research question, for example:

- Is there a statistically significant difference between two groups?
- Is there a statistically significant relationship between two numerical variables?

Task 2 - Hypotheses formulation (5 points)

For the chosen research question, clearly define:

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)

Hypotheses must be written in correct statistical form.

Task 3 - Test selection & justification (5 points)

1. Choose an appropriate statistical test (for example: t-test, correlation test).

2. Briefly justify your choice by explaining:

- the types of variables involved,
- the number of groups or samples,
- why this test is suitable.

Formal assumption testing is not required.

Task 4 - Hypothesis testing in Python (10 points)

Using Python libraries (NumPy, Pandas, SciPy):

1. Perform the selected statistical test.

2. Report:

- test statistic,
- p-value,
- significance level ($\alpha = 0.05$).

The code must be clear, readable, and reproducible.

Task 5 - Interpretation of results (5 points)

Interpret the results in words:

- compare the p-value with the significance level,

- state whether the null hypothesis is rejected or not,
- explain what the result means in the context of the dataset.

Task 6 - Conclusion & limitations (5 points)

Write a short conclusion answering:

- What did you learn from this hypothesis test?
- What does the result imply about the data?
- What are the limitations of this analysis?
- What could be explored further?

Submission requirements

Students must submit ONE ZIP or folder named as *Name_Surname_Assignment2.zip.* containing:

1. Jupyter Notebook (.ipynb)
   - all Python code,
   - outputs,
   - brief comments explaining steps.
2. Report (PDF-formatted)
   - dataset description,
   - research question and hypotheses,
   - explanation of the chosen test,
   - interpretation of results,
   - conclusion and limitations.

Academic integrity note

You may use AI tools (e.g., ChatGPT) for clarification or understanding of concepts. However, all analysis, reasoning, and interpretations must reflect your own understanding. If you cannot explain your work, it may be considered an academic integrity issue. ***AI-generated reports may result in a grade reduction***