Project 12 - DAISY

Student Project Proposal

DAISY – Data Analysis, Inference and Statistical studY

Master area	AI for Drug Discovery; Statistics; Inference
Project type	Computational project with Python implementation, statistical data analysis
Duration	1-2 months
Supervision	One tutor or co-tutor; weekly or bi-weekly progress meetings
Recommended level	Basic knowledge of Python language and packages used for statistical computation; basic knowledge of Statistics and Probability are useful
Main output	Documented code, a technical report summarizing the main deliverables, and a final presentation of the results and conclusions
Possible datasets	Public or tutor-provided data

Short description. The aim of the project is to conduct a full statistical data analysis pipeline over a real dataset, in accordance with the typical workflow used in data science applications. Through the analysis, the students will perform inference about the data and draw statistically sound and robust conclusions on the observed phenomena.

1. Background and rationale

Statistical data analysis provides a systematic framework for extracting meaningful information from observed data, with the aim of supporting scientific investigation and evidence-based decision-making. Statistical methods are typically applied to a sample rather than to the entire population of the study, therefore requiring to properly account for the uncertainty associated with sampling that affects the resulting estimates and decisions.

In many practical applications, statistical analyses can be naturally organized into three complementary stages: *i)* data examination, which provides an initial understanding of the dataset and allows the identification of potential relationships among the observed variables; *ii)* statistical inference, which relates to the estimation of unknown population characteristics while explicitly accounting for the uncertainty of the observations; *iii)* statistical decision-making, where hypothesis tests assess whether the available evidence supports specific claims about the population and where statistical methods can be used to evaluate the performance and effectiveness of simple decision rules.

This project is designed to guide students through each of these stages using a real-world dataset. By completing the proposed activities, students will gain practical experience in the complete statistical workflow, from data exploration to inference and decision analysis, while developing the ability to critically interpret statistical results and their practical implications.

2. Project objective

The objective of the project is to apply statistical data analysis techniques to a real-world dataset through a workflow based on descriptive statistics, statistical inference, and statistical decision analysis. Students will translate the theoretical concepts presented during the course into a complete practical analysis.

The work should be carried out by focusing on the methodological fundamentals discussed during the course and producing simple, well-structured, and well-documented software implementations, with clear and interpretable outputs. The results obtained should be critically analyzed and discussed considering the characteristics of the adopted dataset.

3. Specific aims

- Perform preliminary data examination and preparation by identifying the variables' types, handling missing or inconsistent values, and producing a cleaned dataset that can be processed through statistical methods.
- Identify summary properties describing the data using descriptive statistics, graphical representations, distribution analysis, and correlation analysis.
- Estimate unknown population parameters using sample estimators and confidence intervals.
- Define and evaluate hypotheses testing problems about the data, interpreting p-values and significance levels, and evaluate decision errors and diagnostic performance with appropriate metrics.
- Develop reproducible statistical analyses by implementing well-structured software solutions, presenting results through appropriate tables and graphical representations, and critically interpreting the obtained findings.

4. Proposed methodology

4.1 Data examination

The students start with a crucial preliminary stage by examining the available dataset. During this stage, the students should identify the variables and their type, together with the data quality (e.g., check for missing, inconsistent, or non-admissible values). The goal is to properly preprocess the data to obtain a clean and documented dataset suitable for further statistical analyses.

4.2 Data description

The cleaned available dataset is then summarized through the application of descriptive statistical methods to derive measures of central tendency of the variables, variability indices, empirical distributions, and correlation analysis to detect possible relationships among the variables.

4.3 Parameter estimation

By employing appropriate sample estimators, some selected unknown population parameters associated with the dataset variables are estimated. The uncertainty associated with these estimates is quantified by means of confidence intervals.

4.4 Hypothesis testing and decision analysis

Hypothesis tests for selected population parameters are defined and evaluated, using appropriate significance levels. In particular, different families of tests are taken into account during the project development. Simple decision rules are then evaluated through decision errors and diagnostic performance metrics used in decision theory.

5. Indicative work plan

Period	Main task	Expected output
Week 1	Data examination	Preliminary understanding and examination, reported with a short description, of the dataset and the problem associated with it, data cleaning for successive analyses.
Weeks 2	Data description	Computation of descriptive statistics for the dataset, properly collected and discussed through a short report.
Weeks 3-4	Parameter estimation	Notebooks/scripts to perform sample estimation of unknown parameters, with appropriate confidence intervals derivation.
Weeks 5-6	Hypothesis testing and decision analysis	Formulation of hypothesis tests and decision problems. Notebooks/scripts to evaluate the significance of the tests and the decision performance.
Weeks 7	Integration and reporting	Collection and integration of the results and deliverables (notebooks/scripts, tables, plots, figures), preparation of a final overall report and of a short presentation.

6. Expected deliverables

- Well-structured and well-documented Python notebooks/scripts implementing the statistical data analysis workflow.
- Suitable tables, graphs, and plots useful to corroborate the performed analysis.
- A technical report of approximately 6-8 pages that summarizes the entire workflow, including aims, methods, results, and critical interpretations of the obtained results.
- A final presentation of approximately 10-15 minutes.

7. Skills acquired by the student

- Handling of real data using popular libraries and technologies.
- Capability of interpreting, through descriptive statistics, the behavior of random variables.
- Implementation and validation of parameter estimation routines with uncertainty quantification.
- Implementation and validation of hypothesis tests and decision problems with uncertainty quantification.
- Critical and interpretable evaluation of the results from statistical inference methodologies applied to real data.

8. Suggested software and resources

The whole project can be implemented in Python using standard scientific libraries, including, but not limited to: NumPy, SciPy, Scikit-Learn, Pandas, matplotlib for visualization.

The students can propose a dataset if they prefer. Otherwise, the tutor will provide the dataset to examine. The detailed requirements and tasks to carry out during the project (e.g., the specific parameters to estimate, the statistical tests to conduct, etc.) will be arranged in accordance by the tutor and the students when the project is assigned.

9. Evaluation criteria

Criterion	What should be assessed
Scientific clarity	The statistical data analysis workflow is correct, and the objectives and methodologies applied across the project are clear.
Technical implementation	The code is well-structured, well-documented and easy to read, and reproducible on the selected dataset.
AI relevance	The project deals with fundamental statistical concepts and methodologies that can be applied to any AI-based system.
Interpretability	The students are aware of the analysis outcomes and can interpret the results correctly and critically.
Critical discussion	The final report and presentation discuss the results, limits, uncertainty arising during the project development, and possible future works identified by the students.

10. Optional extensions

For students with stronger computational skills, the project may be extended by adding a regression analysis or logistic regression analysis over selected variables of the dataset.

Note for tutors. The project should be adapted to the available dataset and to the student background. The minimal successful outcome is a clear, reproducible prototype; advanced machine learning components should be treated as optional extensions rather than mandatory requirements.