Inferential Statistics Project

Weighting of this Project and Deadlines

This component will account for the 70% of the final grade. After the submission of the report, there will be a short viva (15 minutes per student) in which we will discuss your project with you. The viva will be used to further assess your understanding of your report. The questions that will be asked in the viva will focus on the material you produced (e.g., explain why you chose this specific hypothesis test, could you have done other types of tests, what is the meaning of this regression coefficient). The weighting of the viva will be 40% of the final project.

- The deadline to submit the project is on **Monday, 17/3/2025 at 16:00 (London time)**.
- The viva will be held on the same week of the deadline.

Project Brief

Objective. You will undertake an individual four-week project to apply and demonstrate core skills in inferential statistics. You must select **one dataset** from the ones provided (each dataset has different characteristics and features). You can also pick a dataset on your own, if you have a specific preference but you will need to ask the lecturer or the GTAs for approval.

Using your chosen dataset, you will conduct an end-to-end exploratory and inferential analysis, culminating in a short written report.

Deliverables.

- **Short Report (At most 7 Pages)** in PDF format describing:
  - Brief presentation of the dataset.
  - Exploratory plots, summary statistics.
  - Verification of central limit theorem (CLT) aspects using the data.
  - Hypothesis testing procedures and results.
  - Correlation and regression analysis.
  - Conclusions (optional).
- **Code/Notebook** (e.g. Python, R, or another language) used for your analysis.

Project Tasks

Below are the *minimum required* tasks. You may go beyond these if it helps highlight interesting aspects of your dataset.

1. Data Exploration (15 Points)

- **Plot distributions** of at least three numeric features (e.g. histograms, density plots).
- **Compute mean and variance** for the corresponding numeric features (and any additional summary statistics if desired).
- Identify any outliers in the distributions that lie outside three standard deviations from the mean.

2. Verifying the Central Limit Theorem (CLT) (20 Points)

- Choose a numeric variable (with enough data) and **draw repeated samples** of various sizes (e.g. 10, 50, 200).
- For each sample size, compute the **sample means** repeatedly (like a resampling approach).
- Visualize how these sample means **distribute**, and discuss how it **relates to the Central Limit Theorem and the Law of Large Numbers**.

3. Hypothesis Testing (25 Points)

- Perform at least one **hypothesis test** relevant to your dataset. For example (not exhaustive, you can come up with different ones):
  - Test whether the mean of a numeric feature equals a hypothesized value.
  - Compare means across two groups.
- **State the null and alternative hypotheses** clearly, choose a significance level (e.g. $\alpha = 0.05$), show the test statistic or p-value, and interpret the result.
- Calculate confidence interval (you can pick the boundary).

4. Correlation Between Variables (15 Points)

- Identify at least three couples of numeric variables in your dataset and **compute their correlations** (You can pick variables A B C and see how they correlate to variable D).
- Create a scatter plot, for each couple of variables, and interpret the strength/direction of the relationship.

5. Regression Analysis and CLT Consideration (25 Points)

- Choose a numeric variable to be the **response (dependent)** variable, and at least one other variable (or multiple) as the **predictor(s)**.
- Fit a **linear regression model** (simple or multiple). Summarize the **regression coefficients**, check the model fit (e.g. R^2), and interpret your findings.
- To illustrate how the **CLT** impacts estimates of standard error in regression, **randomly sample subsets** of your dataset multiple times, re-fit the model each time, and observe how the estimated coefficients (and their standard errors) vary. Briefly discuss how these variations align with the Central Limit Theorem.

Assessment Criteria

Your project will be graded on:

- **Correctness and clarity** of statistical methods (plots, calculations, tests).
- **Depth of interpretation** for results (CLT observations, hypothesis test conclusions, regression findings).
- **Quality of communication** in the written report (coherence, good structure, appropriate use of visuals).
- **Reproducibility** of your code (clearly labeled and well-documented).

Total Points: 100

Submission Instructions

- Submit your PDF and the code via Blackboard by the deadline indicated at the top of this document.
- Ensure that your student ID is included on the first page of your document. Do not add your name to the PDF.
- Make sure that you submit your work in advance of the time of the deadline. Late submissions will be penalized according to the college late submission policy, which can be found here. The policy states that any work submitted up to one (1) day after the assessment deadline (date and time) will be marked but capped at the passmark. Work submitted more than one (1) day late will not be accepted as a valid attempt and mark of zero will be recorded.

Academic Integrity

Remember to adhere to Imperial Code of Conduct, available here. All work must be your own, and any sources used must be properly cited.

Good Luck!
