Coursework Mathematical Statistics
Academic Year: 2025-2026

Instructions

- Please choose and solve **one** of the problems listed below. You must submit:
  1. The **analytical solution**: **hand-written, attached as a (scanned) pdf**. Please scan it properly - unclear pictures or draft versions will not be taken into account.
  2. The **Python code**: **please provide a link to a Google Colab notebook where you write the code**. The notebook should be set up such that I have access to view it and to run the corresponding code. Display/plot the most relevant quantities for the problem you chose, and write a short report (2–3 paragraphs - these can be added in a text cell in the notebook created by you) discussing how and why the outputs change when selected input parameters are modified. **Choose the parameters you consider most relevant for the chosen problem**.
- **Submit your complete work (explanations, solutions, and code) via Teams. E.g. You can attach a pdf with the analytical solution and a link to the notebook.** All submitted solutions must represent the student's own independent work. Solutions or parts of solutions generated using generative artificial intelligence tools (such as ChatGPT) will not be accepted and will not be considered for assessment.
- **Deadline: 25th of January 2026.**

Problem 1

Daniel claims to Emma that his average time across six cycling trials is lower, and therefore he is the faster cyclist. Emma disagrees and argues that she is more reliable because the variability of her times is smaller. The recorded completion times (assumed independent and normally distributed) are shown below.

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
Daniel	18.4	17.9	18.1	19.2	18.6	17.7
Emma	18.9	18.7	18.8	18.4	18.6	18.5

(a) Is there sufficient statistical evidence to support Daniel's claim? Clearly state the null hypothesis H_0 and the alternative hypothesis H_A. Test whether the population variances are equal and select an appropriate two-sample t-test. Perform the test and state your conclusion.

(b) Is there sufficient statistical evidence to support Emma's claim? Formulate the hypotheses H_0 and H_A and carry out the appropriate test.

Use a significance level of 5% throughout.

Problem 2

Part A. A biologist investigates the average concentration of a particular enzyme in blood samples. Previous studies suggest the concentrations follow a Normal distribution with known standard deviation 0.15 mg/L. A random sample of 64 individuals yields an average concentration of 1.32 mg/L.

(a) Construct a 95% confidence interval for the population mean enzyme concentration.

(b) Assuming the true population mean concentration is 1.28 mg/L, compute the probability that the sample mean from 64 individuals is at least 1.32 mg/L.

Part B. A survey organization reports results from two regions. It claims that support for a policy proposal is higher in region X than in region Y by 12 percentage points. The reported support levels are 58% in region X and 46% in region Y. Each region was sampled independently with 1,000 randomly selected respondents. Determine the margin of error corresponding to the 95% confidence interval for each estimate: 58%, 46%, and the difference of 12%.

Problem 3

Environmental researchers study the relationship between the age of residential buildings (measured in years since construction) and their annual energy consumption (measured in megawatt-hours, MWh). A sample of 200 buildings was analyzed, with summary statistics given below:

	Sample mean	Standard deviation
Building age (years)	31.4	18.2
Energy consumption (MWh)	14.7	4.1

The sample correlation coefficient is $r = 0.21$.

(a) Derive the least squares regression line predicting annual energy consumption based on building age. Interpret both the slope and the intercept.

(b) Compute and interpret R^2. Comment on whether this model provides a good fit.

(c) Predict the annual energy consumption of a building that is 40 years old. Construct a 95% prediction interval for a single building and a 95% confidence interval for the mean energy consumption of all buildings of this age.

Problem 4

A data scientist examines whether the number of hours spent tuning a predictive algorithm affects its classification accuracy. Data from 12 tuning sessions record the tuning time (in hours) and the resulting accuracy (as a percentage).

The mean tuning time is 5.1 hours with standard deviation 0.8 hours. The mean accuracy is 76% with standard deviation 11%. The sample correlation between tuning time and accuracy is $r = 0.55$.

(a) Determine the regression equation for predicting accuracy based on tuning time.

(b) Predict the accuracy if the algorithm is tuned for 6 hours.

(c) Does this linear regression model explain a substantial proportion of the variability in accuracy? Provide a statistical justification.