Statistics & Probability

Course Project

Analyzing Factors That Influence
Student Academic Performance

Dataset: UCI Student Performance (Mathematics)
Source data file student-mat.csv (inside Zip file): archive.ics.uci.edu/ml/datasets/Student+Performance

Also available on Kaggle (student-mat.csv)

1. Project Overview

Use Colab Notebook to write Python code to apply core concepts from Statistics and Probability course to a real-world dataset. You will perform a complete statistical analysis of factors that influence student academic performance in mathematics, using the UCI Student Performance dataset.

Learning Objectives

Compute and interpret descriptive statistics (mean, median, mode, variance).
Construct meaningful visualizations (histograms, boxplots, scatter plots).
Apply probability rules including conditional probability and Bayes' Theorem.
Formulate and execute hypothesis tests.
Perform correlation analysis.
Construct and interpret confidence intervals.
Communicate statistical findings clearly in a written report.

Tools & Libraries

Python 3.8+
pandas — data manipulation
numpy — numerical computation
matplotlib / seaborn — visualization
scipy.stats — statistical tests and distributions

2. Dataset Description

The dataset comes from the UCI Machine Learning Repository and contains student achievement data from two Portuguese schools. This project uses the Mathematics course file (`student-mat.csv`). The dataset has **395 student records** and **33 attributes**.

Key Variables Used in This Project

Variable	Type	Description
G1	Numeric (0–20)	First period grade
G2	Numeric (0–20)	Second period grade
G3	Numeric (0–20)	Final grade (target variable)
studytime	Ordinal (1–4)	Weekly study time: 1 = <2h, 2 = 2–5h, 3 = 5–10h, 4 = >10h
sex	Binary (M/F)	Student's sex
internet	Binary (yes/no)	Internet access at home
failures	Numeric (0–4)	Number of past class failures
absences	Numeric (0–93)	Number of school absences
romantic	Binary (yes/no)	Whether student is in a romantic relationship
Mjob	Categorical	Mother's job (teacher, health, services, at_home, other)
age	Numeric (15–22)	Student's age
pass	Binary (0/1)	Derived: 1 if G3 ≥ 10, else 0

The binary variable `pass` is computed from G3: a student passes if their final grade is 10 or above (out of 20).

3. Descriptive Statistics

In this section you will compute and interpret summary statistics for key numeric variables and create frequency tables for categorical variables.

- Compute the following for the final grade (G3): mean, median, mode, standard deviation, variance, and IQR. Also produce a frequency table for studytime.
- Compute the same descriptive measures (mean, median, mode, std dev, variance, IQR) for G1 and G2. Compare them with G3.

4. Data Visualization & Distributions

Visualization is critical for understanding data distributions, spotting outliers, and checking assumptions required by statistical tests.

Visualizations

- Create histogram of G3: Show the distribution of final grades with mean and median lines.
- Create Boxplot of G3 by Sex: Compare grade distributions between male and female students.
- Create Plot Bar Chart of Pass/Fail by Internet Access: Display counts of pass vs. fail for students with and without internet.
- Draw scatter Plot of G1 vs G3: Explores the relationship between first-period and final grades.
- Create a histogram of the `absences` variable. Describe its shape (symmetric, skewed left/right, unimodal/multimodal, any outliers).
- Determine whether the final grade G3 is approximately normally distributed.

5. Probability & Conditional Probability

You will compute the marginal, joint, and conditional probabilities — as well as Bayes' Theorem to real data.

- Compute Marginal probabilities: P(Pass) and P(Internet = yes).

- Compute Joint probability: P(Pass AND Internet = yes).
- Compute Conditional probabilities: P(Pass | Internet = yes) and P(Pass | Internet = no).
- Bayes' Theorem: P(Internet = yes | Pass), verified by direct computation.
- Compute P(Pass | studytime ≥ 3) and P(Pass | studytime < 3). Compare the two probabilities and discuss whether more study time appears to help.
- Create a full joint probability table (using `pd.crosstab` with `normalize='all'`) for `internet` vs `pass`. Include marginal totals (`margins=True`). Present the table and briefly interpret it.

6. Hypothesis Testing

For every test, you must follow the standard five-step procedure:

1. **Step 1:** State the null hypothesis (H_0) and alternative hypothesis (H_1).
2. **Step 2:** Choose the significance level ($\alpha = 0.05$ unless stated otherwise).
3. **Step 3:** Check assumptions and select the appropriate test.
4. **Step 4:** Compute the test statistic and p-value.
5. **Step 5:** State your decision (reject or fail to reject H_0) and interpret in context.

Test 1: One-Sample t-Test

Research question: Is the average final grade equal to 10 (the pass mark)?

- $H_0 : \mu = 10$
- $H_1 : \mu \neq 10$

Your code should print the t-statistic, p-value, and decision.

Test 2: One-Sample z-Test

Research question: Is the passing rate for students greater than 50%?

- $H_0 : p <= 50\%$
- $H_1 : p > 50\%$

Your code should print the t-statistic, p-value, and decision.

7. Correlation coefficient

This section explores linear relationships between variables using correlation coefficients.

- Find correlation coefficient between studytime and G3
- Find correlation coefficient between G1 and G3

8. Confidence Intervals

Confidence intervals provide a range of plausible values for a population parameter and complement hypothesis testing.

- Compute 95% confidence interval for the population mean final grade.
- Compute a 99% confidence interval for the mean final grade. Compare it with the 95% CI. Explain why the 99% interval is wider and what this means in terms of the trade-off between confidence and precision.
- Compute separate 95% confidence intervals for the mean final grade of students WITH internet access and students WITHOUT internet access. Do these intervals overlap? Discuss what the overlap (or lack thereof) suggests about the difference between the two groups.

9. Final Report Requirements

In addition to the completed code, submit a written summary report (1–2 pages) covering the following sections:

Dataset Description: How many students? What variables did you analyze? Any data quality issues?

Key Descriptive Findings: What is the typical student grade? How spread out are grades? Which variables showed interesting patterns?

Probability Insights: What is the overall pass rate? Does internet access change the probability of passing?

Hypothesis Testing Results: Summarize hypotheses test: Statistic, p-value, Decision at $\alpha = 0.05$.

Confidence Intervals: Interpret the CI for the mean grade. What happens when you move from 95% to 99%?

Limitations & Future Work: What are the limitations of this analysis? What additional tests or methods would you try next?