Student Performance & Study Habits: A Comprehensive Analysis Using Descriptive and Inferential Statistics

1.2 Goal of the Project

To analyse how students' study habits, attendance, demographics, and learning methods influence their academic performance (midterm and final exam scores) and pass/fail outcomes.

1.3 Why This Project Works

This dataset allows the application of all major concepts in statistics:

- Descriptive Statistics
- Sampling (Simple + Stratified)
- Central Limit Theorem
- Measures of central tendency and dispersion
- Outlier detection (IQR + Z-score)
- Correlation & Regression
- Hypothesis Testing: t-test, ANOVA, Chi-square
- Multiple Testing (Bonferroni Correction)

2. Dataset Description

Column	Type	Description
student_id	ID	Unique identifier
school	Categorical	School A/B/C
gender	Categorical	Male/Female
study_hours_week	Numeric	Hours studied per week
study_method	Categorical	Solo/Group/Tutor

Column	Type	Description
attendance_rate	Numeric	Attendance percentage
midterm_score	Numeric	Midterm exam score
final_score	Numeric	Final exam score
passed_final	Categorical	Pass/Fail outcome
socioeconomic_idx	Numeric	SES factor

Dataset: 300 students

Variables: 10 (1 ID, 5 numeric, 4 categorical)

Missing Values: None

Time to Load: < 1 second

Memory: ~150 KB

3. Sampling & Theoretical Foundations

A. Simple Random Sampling

- **Purpose:** Ensure unbiased representation
- **Sample Size:** 50 students
- **Method:** sample(n=50, replace=False)
- Population Mean (all 300): final_score = 36.34
- Sample Mean (n=50): final_score = 36.89
- Difference (sampling error): +0.55 (1.5% error)
- Population SD: 13.05
- Sample SD: 12.87
- Standard Error (SE): $13.05 / \sqrt{50} = 1.85$

B. Stratified Sampling

- **Purpose:** Ensure representation across subgroups
- **Strata: School × Gender (6 combinations)**

School A-Male:	10 samples
School A-Female:	10 samples
School B-Male:	10 samples
School B-Female:	10 samples
School C-Male:	10 samples
School C-Female:	10 samples

Total: 60 samples

- **Method:** groupby('school', 'gender') → sample from each group

4. Central Limit Theorem (CLT) Demonstration

1. Create a population distribution of final scores
2. Draw multiple random samples of size n (n = 5, 30, 100)
3. Calculate the mean of each sample (repeat 300 times)
4. Plot the distribution of sample means
5. Observe convergence to normality

Property	**n=5**	**n=30**	**n=100**	**Trend**
Distribution Shape	Non-smooth	Bell curve	Perfect bell	Converges to normal
Skewness	0.24	0.08	0.02	Approaches 0
SE Accuracy	101%	100%	100%	Theory matches
Concentration	Spread	Moderate	Tight	Tighter around mean

5. Descriptive Statistics

Measures to Calculate:

Central Tendency

- **Mean (μ):** Average value
- **Median (M):** Middle value when sorted
- **Mode:** Most frequent value

Dispersion

- **Standard Deviation (σ):** Average spread from mean
- **Range:** Max - Min
- **IQR (Interquartile Range):** Q3 - Q1

Shape

- **Skewness:** Asymmetry (-0.5 to +0.5 = roughly symmetric)
- **Kurtosis:** Tail heaviness (positive = heavy tails, negative = light tails)

Variable	Mean	Median	Std Dev	Min	Max	Range	IQR	Skew	Kurt
study_hours_week	12.33	12.43	5.27	0.06	28.27	28.21	6.51	0.12	-0.14
attendance_rate	85.33	85.49	9.65	43.53	100.00	56.47	13.33	-0.34	-0.47
midterm_score	31.31	31.72	10.82	0.00	58.53	58.53	13.91	-0.07	-0.13
final_score	36.34	37.73	13.05	0.00	65.25	65.25	17.67	-0.18	-0.37

Statistical Interpretations

Study Hours Per Week

- **Mean = 12.33 hours:** Students study ~2.4 hours daily on average
- **SD = 5.27:** High variability; some students study minimal hours, others 25+
- **Skew = 0.12:** Roughly symmetric distribution
- **Insight:** Wide range suggests heterogeneous study habits

Attendance Rate

- **Mean = 85.33%:** Good average attendance across cohort
- **Skew = -0.34:** Left-skewed; most students have high attendance
- **Kurtosis = -0.47:** Flatter distribution (platykurtic); light tails
- **Insight:** Attendance is consistently high; less variable than study hours

Midterm vs Final Scores

- **Midterm Mean = 31.31**, Final Mean = 36.34 (+5.03 point improvement)
- **Final SD = 13.05 > Midterm SD = 10.82:** More spread in final scores
- **Both negatively skewed:** Concentration at lower-middle range
- **Insight:** Students show improvement but both distributions are left-skewed

6. Outlier Detection

Using IQR Method:

Lower Bound = Q1 - 1.5 × IQR

Upper Bound = Q3 + 1.5 × IQR

Outliers: Values outside these bounds

- study_hours_week → **2 outliers**(Row IDs: [247, 289] Values: [26.12, 27.95])
- attendance_rate → **1**(Value: 43.53)
- midterm_score → **2**(Row IDs: [112, 245] Values: [58.12, 58.53])
- final_score → **0**

Using Z-score Method (|z| > 3):

z = (x - mean) / SD

Outliers: |z| > 3 (extremely rare values)

- study_hours_week → **1**(z = 3.12)

Conclusion:
Dataset is mostly clean; very few extreme values.

7. Hypothesis Testing

Hypothesis A – Two-Sample Welch's T-Test

Research Question:
Do students who study more than 10 hours/week score higher?

- High-study mean = **37.49**
- Low-study mean = **34.04**
- t = **2.1423**, p = **0.0167**
- Effect size (Cohen's d) = **0.266** (small)

Decision: Reject H_0
Interpretation:
High-study students score significantly higher — but effect size is **small**.

Hypothesis B – Paired T-test

Research Question:
Did students improve from midterm to final exam?

- t = **11.20**, p < **0.0001**

Decision: Reject H_0
Interpretation:
There is a **significant improvement** from midterm to final exam.

Hypothesis C – ANOVA

Research Question:
Do different schools differ in final scores?

- F = **0.3978**, p = **0.6722**
- Effect size η^2 = **0.0027** (negligible)

Decision: Fail to Reject H_0
Interpretation:
Final scores **do not differ** based on school.

Hypothesis D – Chi-Square Test

Research Question:
Is study method related to pass/fail outcome?

Contingency table showed:

- χ^2 = **1.5542**, p = **0.4597**

Decision: Fail to Reject H_0
Interpretation:
Study method (Solo/Group/Tutor) does **not** influence pass/fail rate.

Hypothesis E – Correlation Analysis

Pearson r = 0.1533, p = **0.0078**

Interpretation:
There is a **weak positive correlation** between study hours and final score.

8. Bonferroni Multiple Testing Correction

Test	**p-value**	**Significant (α=0.05)**	**Decision**
Welch's t-test	0.0167	Yes	**Not significant after correction**
Paired t-test	<0.0001	Yes	**Still significant**
ANOVA	0.6722	No	**Remains non-significant**
Chi-square	0.4597	No	**Remains non-significant**
Correlation	0.0078	Yes	**Still significant**

Conclusion After Correction

Only:

- **Paired T-test**, and
- **Correlation**

remain statistically significant after adjustment.

9. Final Interpretation & Discussion

Key Findings

1. **More study hours → Slight improvement** in final scores.
2. **Students significantly improve** from the midterm to the final exam.
3. **School does not impact performance**.
4. **Study method does not affect pass rate**.
5. A weak but **significant correlation** exists between study hours and performance.

Practical Implications

- Improving student habits (study time, consistency) can increase scores.
- Structured learning programs may not impact as much as self-driven study hours.
- Attendance is consistently high and stable—less of a predictor.

10. Conclusion

This project successfully demonstrates:

- Complete descriptive and inferential statistical workflow
- Clean sampling strategy
- Valid verification of the Central Limit Theorem
- Accurate hypothesis testing
- Proper use of effect sizes and corrections

The analysis provides meaningful insights into how study behaviors affect academic outcomes and supports evidence-based decision-making for student performance improvement

Data Quality Assessment

Missing Values: 0/300 (perfect)

Outliers (extreme): 1/300 (0.33% - minimal)

Data Integrity: Excellent

Assumptions Met: Mostly(minor deviations acceptable)

Sample Size:n=300 (robust, sufficient power)

Overall Quality: High