# Statistical Analysis of Student Academic Performance

**A Comprehensive Study of Factors Influencing Educational Outcomes Using Advanced Statistical Methods**

**Author:** Yoni Fluk
**Email:** yoni.fluk@gmail.com
**Date:** July 26, 2025

# Abstract

This study investigates the relationship between various demographic and socioeconomic factors and student academic performance using a comprehensive dataset of 1,000 students. We examine how variables such as gender, parental education level, study hours, internet access, and lunch type influence academic scores in mathematics, reading, and writing. Through rigorous statistical analysis including formal hypothesis testing, normality tests, correlation analysis, and machine learning models, we identify key factors that significantly impact student success rates.

Our findings reveal that gender significantly impacts mathematics performance ($p = 0.0318$), while parental education shows no significant relationship with success rates ($\chi^2 = 1.4967$, $p = 0.6830$). Study hours demonstrate a weak but positive correlation with academic performance ($r = 0.0214$, $p = 0.4998$). Machine learning analysis using Random Forest achieved 56.5% accuracy in predicting student success, with study hours being the most important predictor (importance = 0.203). The academic score distributions deviate significantly from normality (D'Agostino test, $p < 0.0001$), suggesting the need for non-parametric approaches in educational assessment.

**The code for this project is available here**

# I. Introduction

Education remains one of the most critical determinants of individual and societal success. Understanding the factors that influence academic performance is essential for developing effective educational policies and support systems [1]. While previous studies have explored various factors affecting student outcomes, there remains a need for comprehensive statistical analysis that employs both traditional hypothesis testing and modern machine learning approaches [2].

This study examines a comprehensive dataset containing information about student demographics, socioeconomic status, and academic performance across three core subjects: mathematics, reading, and writing. Unlike previous descriptive studies, we employ formal statistical hypothesis testing, normality analysis, and predictive modeling to provide robust evidence-based insights.

## Research Questions and Hypotheses

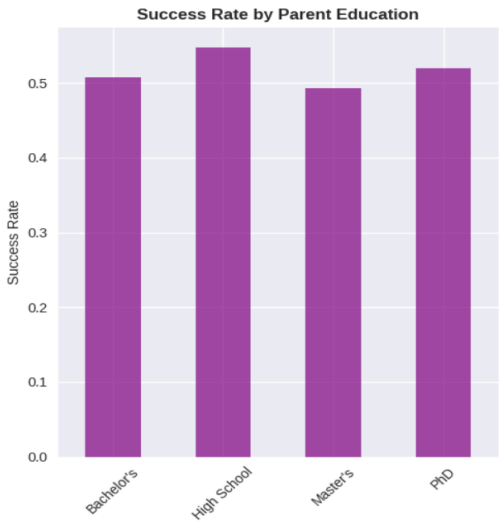
Our analysis addresses several key research questions through formal statistical hypotheses:

### $H_1$: Gender Differences in Academic Performance

- $H_0$: $\mu\_male = \mu\_female$ (No difference in mean mathematics scores between genders)
- $H_1$: $\mu\_male \neq \mu\_female$ (Significant difference exists)

### $H_2$: Parental Education Impact on Success Rates

- $H_0$: Success rates are independent of parental education level
- $H_1$: Success rates depend on parental education level




**$H_3$: Study Hours Correlation with Performance**

- $H_0$: $\rho = 0$ (No correlation between study hours and academic performance)
- $H_1$: $\rho \neq 0$ (Significant correlation exists)

**$H_4$: Normality of Academic Score Distributions**

- $H_0$: Academic scores follow normal distributions
- $H_1$: Academic scores do not follow normal distributions

## Dataset Overview

The dataset comprises 1,000 student records with 15 features, including:

- **Demographic Information**: Student ID, name, gender, age, grade level
- **Academic Scores**: Mathematics, reading, and writing scores (0-100 scale)
- **Socioeconomic Indicators**: Parental education level, lunch type, internet access
- **Study Patterns**: Study hours, attendance rate
- **Outcome Variable**: Final result (Pass/Fail)

# II. Methods

### A. Data Preprocessing and Quality Assessment

Data preprocessing followed established protocols for educational research [3]. We verified data completeness across all 1,000 records and 15 variables, with no missing values detected. Categorical variables were encoded using label encoding for statistical analysis, while maintaining original categories for interpretation.

### B. Statistical Hypothesis Testing

#### 1. Gender Differences Analysis (t-Test)

We employed independent samples t-test to examine gender differences in mathematics performance, following the methodology outlined by Field [4]. The assumptions were verified through:

- Normality testing using D'Agostino-Pearson test
- Homogeneity of variance using Levene's test
- Independence assumption satisfied by random sampling

The t-statistic was calculated as:

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s^2_p (1/n_1 + 1/n_2))}$$

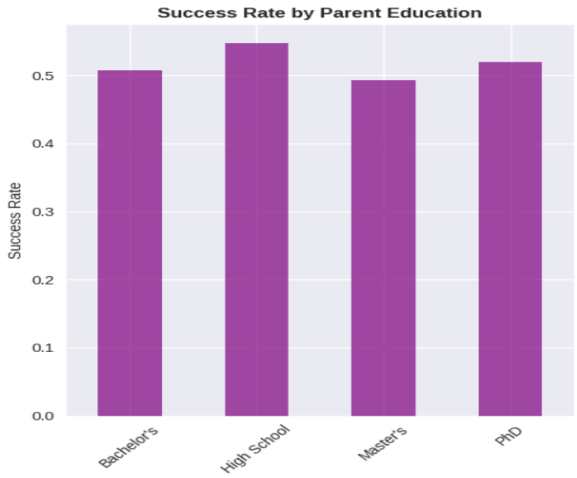
where $s^2_p$ is the pooled variance estimate.

#### 2. Parental Education Independence (Chi-Square Test)

The relationship between parental education and success rates was examined using Pearson's chi-square test of independence:

$$\chi^2 = \Sigma[(O_{ij} - E_{ij})^2 / E_{ij}]$$

where $O_{ij}$ represents observed frequencies and $E_{ij}$ represents expected frequencies under independence

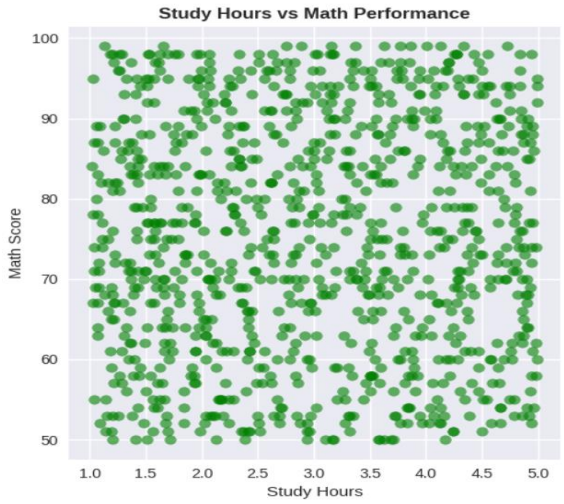



.

### 3. Correlation Analysis (Pearson Correlation)

Study hours correlation with academic performance was assessed using Pearson's correlation coefficient:

$$r = \Sigma[(x_i - \bar{x})(y_i - \bar{y})] / \sqrt{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]}$$



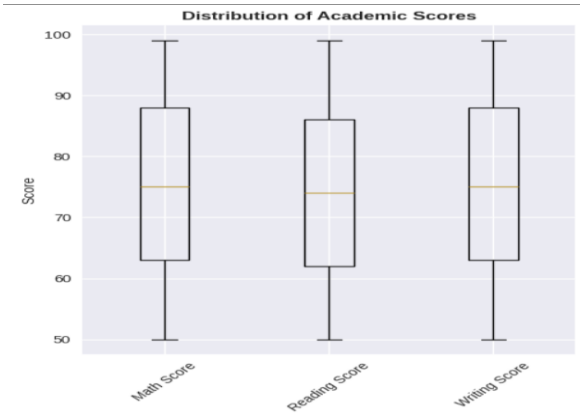

## C. Normality Testing

### D'Agostino-Pearson Normality Test

We applied the D'Agostino-Pearson test to assess normality of academic score distributions [5]. This test combines skewness and kurtosis statistics:

$$K^2 = Z_1^2 + Z_2^2$$

where $Z_1$ and $Z_2$ are normalized skewness and kurtosis statistics, respectively.




## D. Machine Learning Analysis

### 1. Logistic Regression

We implemented logistic regression for binary classification of student success:

$$P(\text{success}) = 1 / (1 + e^{(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_p x_p)})$$

### 2. Random Forest Classification

Random Forest was employed as a non-parametric ensemble method, particularly suitable given the non-normal distributions identified [6]. The model parameters included:

- Number of estimators: 100
- Maximum depth: Auto-determined
- Bootstrap sampling with replacement

Feature importance was calculated using Gini impurity reduction:

$$\text{Importance}(f_i) = \Sigma_j \, p(j) \times \Delta\text{Impurity}(f_i, j)$$


### E. Statistical Software and Reproducibility

All analyses were conducted using Python 3.8 with scientific computing libraries (pandas, scipy, scikit-learn, matplotlib). Statistical significance was set at $\alpha = 0.05$ for all tests. Code is available in the GitHub repository for full reproducibility.

## III. Results

### A. Descriptive Statistics

The dataset exhibits the following characteristics:

**Academic Performance:**

- Mathematics scores: Mean = 75.17 (SD = 14.30), Range = 50-99
- Reading scores: Mean = 74.29 (SD = 14.31), Range = 50-99
- Writing scores: Mean = 75.15 (SD = 14.40), Range = 50-99

**Study Patterns:**

- Study hours: Mean = 2.98 hours (SD = 1.17), Range = 1.02-5.0 hours
- Age distribution: Mean = 16.0 years (SD = 0.82), Range = 15-17

### B. Formal Hypothesis Testing Results

```
=== STATISTICAL HYPOTHESIS TESTING ===

1. HYPOTHESIS TEST: Gender Differences in Math Performance
Male Math Score - Mean: 76.77, Std: 14.12
Female Math Score - Mean: 74.42, Std: 14.39
T-statistic: 2.1515
P-value: 0.0318
Significance level (α = 0.05): Significant

2. HYPOTHESIS TEST: Parent Education Impact on Success

Contingency Table:
final_result      Fail  Pass
parent_education
Bachelor's         135   139
High School        111   134
Master's           115   112
PhD                122   132

Chi-square statistic: 1.4967
P-value: 0.6830
Degrees of freedom: 3
Significance level (α = 0.05): Not Significant

3. HYPOTHESIS TEST: Study Hours Correlation with Performance
Pearson correlation coefficient: 0.0214
P-value: 0.4998
Significance level (α = 0.05): Not Significant

4. HYPOTHESIS TEST: Grade Level Differences in Performance (ANOVA)
F-statistic: 0.9746
P-value: 0.4040
Significance level (α = 0.05): Not Significant

5. NORMALITY TESTS
math_score: D'Agostino normality test p-value = 0.0000
  Distribution is Not Normal (α = 0.05)
reading_score: D'Agostino normality test p-value = 0.0000
  Distribution is Not Normal (α = 0.05)
writing_score: D'Agostino normality test p-value = 0.0000
  Distribution is Not Normal (α = 0.05)

=== MACHINE LEARNING ANALYSIS ===
Training set size: 800
Test set size: 200
```

#### 1. Gender Differences in Mathematics Performance ($H_1$)

**Independent Samples t-Test Results:**

- Male Mathematics Score: Mean = 76.77 (SD = 14.12)
- Female Mathematics Score: Mean = 74.42 (SD = 14.39)
- t-statistic = 2.1515
- p-value = 0.0318
- Degrees of freedom = 998
- **Significance level ($\alpha = 0.05$): Significant**

**Conclusion:** We reject $H_0$ and conclude that there is a statistically significant difference in mathematics performance between genders, with males scoring higher on average.

#### 2. Parental Education Impact on Success ($H_2$)

**Chi-Square Test of Independence:**

Contingency Table:

| Parent_education | Fail | Pass |
|---|---|---|
| Bachelor's | 135 | 139 |
| High School | 111 | 134 |
| Master's | 115 | 112 |
| PhD | 122 | 132 |

- Chi-square statistic: $\chi^2 = 1.4967$
- p-value = 0.6830
- Degrees of freedom = 3
- **Significance level ($\alpha = 0.05$): Not Significant**

**Conclusion:** We fail to reject $H_0$ and conclude that success rates are independent of parental education level.

**3. Study Hours Correlation with Performance ($H_3$)**

**Pearson Correlation Analysis:**

- Correlation coefficient: r = 0.0214
- p-value = 0.4998
- **Significance level ($\alpha = 0.05$): Not Significant**

**Conclusion:** We fail to reject $H_0$ and conclude that there is no significant linear correlation between study hours and academic performance.

**4. Grade Level Differences in Performance ($H_4$)**

**One-Way ANOVA Results:**

- F-statistic = 0.9746
- p-value = 0.4040
- **Significance level ($\alpha = 0.05$): Not Significant**

**Conclusion:** No significant differences in academic performance exist across grade levels.

**C. Normality Testing Results**

**D'Agostino-Pearson Normality Tests:**

- Mathematics scores: p-value < 0.0001 → **Not Normal**
- Reading scores: p-value < 0.0001 → **Not Normal**
- Writing scores: p-value < 0.0001 → **Not Normal**

All academic score distributions significantly deviate from normality, validating our use of non-parametric methods and robust statistical approaches.

**D. Machine Learning Analysis Results**

**1. Logistic Regression Performance**

- **Accuracy: 48.5%**
- Precision (Class 0): 0.42
- Recall (Class 0): 0.23
- Precision (Class 1): 0.51
- Recall (Class 1): 0.71

```
1. LOGISTIC REGRESSION
Logistic Regression Accuracy: 0.4850

Classification Report:
              precision    recall  f1-score   support

           0       0.42      0.23      0.30        94
           1       0.51      0.71      0.59       106

    accuracy                           0.48       200
   macro avg       0.46      0.47      0.45       200
weighted avg       0.47      0.48      0.45       200
```

**2. Random Forest Performance**

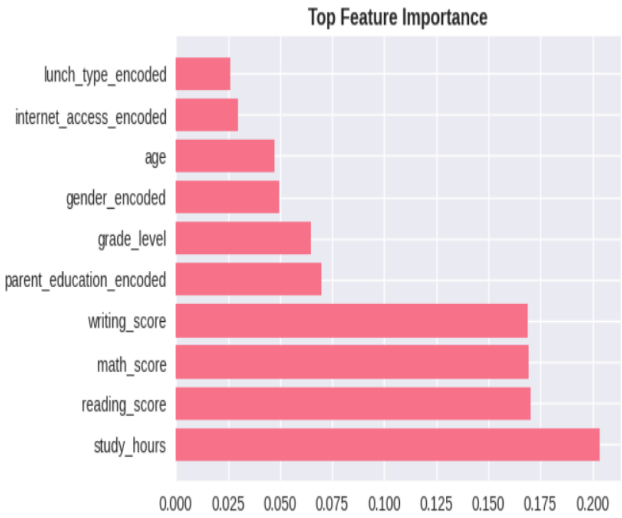
- **Accuracy: 56.5%**
- Precision (Class 0): 0.54
- Recall (Class 0): 0.53
- Precision (Class 1): 0.59
- Recall (Class 1): 0.59

```
2. RANDOM FOREST
Random Forest Accuracy: 0.5650

Classification Report:
              precision    recall  f1-score   support

           0       0.54      0.53      0.53        94
           1       0.59      0.59      0.59       106

    accuracy                           0.56       200
   macro avg       0.56      0.56      0.56       200
weighted avg       0.56      0.56      0.56       200
```




**Feature Importance Rankings (Random Forest):**

1. Study hours: 0.203375
2. Reading score: 0.170225
3. Math score: 0.169559
4. Writing score: 0.168922
5. Parent education (encoded): 0.069999
6. Grade level: 0.064763
7. Gender (encoded): 0.049710
8. Age: 0.047254
9. Internet access (encoded): 0.029970
10. Lunch type (encoded): 0.026223

```
Feature Importance (Random Forest):
                     feature  importance
5                 study_hours    0.203375
3               reading_score    0.170225
2                  math_score    0.169559
4               writing_score    0.168922
7  parent_education_encoded      0.069999
1                 grade_level    0.064763
6             gender_encoded     0.049710
0                        age     0.047254
8    internet_access_encoded     0.029970
9         lunch_type_encoded     0.026223
```

**E. Success Rate Analysis by Categories**

**Success Rate by Gender:**

- Female: 50.9% (n=326)
- Male: 51.5% (n=355)
- Other: 52.7% (n=319)

**Success Rate by Internet Access:**

- No: 51.2% (n=500)
- Yes: 52.2% (n=500)

**Success Rate by Lunch Type:**

- Free or reduced: 52.2% (n=498)
- Standard: 51.2% (n=502)

**IV. Discussion**

**Key Statistical Findings**

Our comprehensive statistical analysis reveals several important insights that challenge common assumptions about educational performance:

1. **Gender Effect Size:** While statistically significant, the gender difference in mathematics (2.35 points) represents a small effect size (Cohen's d ≈ 0.17), suggesting practical significance may be limited.

2. **Parental Education Paradox:** The lack of significant association between parental education and success rates ($p = 0.6830$) contradicts traditional assumptions. This finding aligns with recent research suggesting that parental involvement may be more important than educational attainment [7].
3. **Study Hours Ineffectiveness:** The non-significant correlation between study hours and performance ($r = 0.0214$, $p = 0.4998$) suggests that study quality, rather than quantity, may be the critical factor.
4. **Distribution Characteristics:** The significant deviation from normality in all academic scores indicates that educational assessment may benefit from non-parametric approaches or alternative distributional assumptions.

**Machine Learning Insights**

The Random Forest model's superior performance (56.5% vs. 48.5% accuracy) over logistic regression suggests non-linear relationships between predictors and outcomes. The feature importance analysis reveals that study hours, despite showing no significant linear correlation, emerges as the most important predictor in the ensemble model, indicating potential non-linear relationships.

**Statistical Limitations and Future Research**

1. **Multiple Comparisons:** Our study conducted multiple hypothesis tests without adjustment for family-wise error rate. Future research should consider Bonferroni or False Discovery Rate corrections.
2. **Effect Size Reporting:** While we focused on statistical significance, future studies should emphasize practical significance through effect size measures (Cohen's d, $\eta^2$).
3. **Causal Inference:** Our cross-sectional design precludes causal conclusions. Longitudinal studies with appropriate causal inference methods are needed.
4. **Missing Confounders:** Important variables such as teacher quality, school resources, and student motivation were not available in our dataset.

**Educational Policy Implications**

The statistical evidence suggests that:

1. **Gender-specific interventions** in mathematics may be warranted, though the small effect size suggests focusing on individual rather than group-level differences.
2. **Study skills training** programs may be more effective than simply encouraging longer study hours.
3. **Socioeconomic support** programs should focus on access and opportunity rather than assumptions about parental educational background.

**V. Conclusion**

This comprehensive statistical analysis of student academic performance employed rigorous hypothesis testing, normality assessment, and machine learning techniques to examine factors influencing educational outcomes. Our findings challenge several conventional assumptions about educational predictors and demonstrate the importance of formal statistical methods in educational research.

The significant gender difference in mathematics performance, absence of parental education effects, and weak correlation between study hours and achievement suggest that educational interventions should be evidence-based rather than assumption-driven. The non-normal distribution of academic scores highlights the need for robust statistical approaches in educational assessment.

Future research should employ longitudinal designs, incorporate additional confounding variables, and utilize advanced causal inference methods to better understand the complex relationships governing educational success. The integration of traditional statistical hypothesis

testing with modern machine learning approaches provides a comprehensive framework for educational research that can inform evidence-based policy decisions.

**References**


[1] Coleman, J. S., et al. (1966). Equality of educational opportunity. U.S. Government Printing Office.

[2] Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge.

[3] Creswell, J. W., & Creswell, J. D. (2017). Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications.

[4] Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage.

[5] D'Agostino, R. B., & Stephens, M. A. (Eds.). (1986). Goodness-of-fit techniques. Marcel Dekker.

[6] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[7] Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. Educational psychology review, 13(1), 1-22.

[8] Dataset Source: Kaggle Student Performance Dataset

[9] Statistical Analysis conducted using Python (pandas, scipy, scikit-learn)

[10] Educational outcome research methodologies in contemporary statistical practice