

0. Scenario

I am a statistician teacher at a university tasked with elaborating a case study demonstrates a full end-to-end data analysis workflow focused on hypothesis testing and correlation analysis. The objective was to determine whether weekly self-study time has a statistically significant impact on academic performance.

1. Ask

Is there a statistically significant relationship between weekly self-study hours and academic performance?

2. Prepare

The dataset combines results from a university-run survey with students' performance scores retrieved from the learning registry system. The first step is to perform a ROCCC assessment of the dataset:

- **R – Reliable:** The data is collected entirely by the university using its IT systems and a well-designed survey, so it is considered reliable
- **O – Original:** The data is collected entirely by the university, so it is considered original.
- **C – Comprehensive:** The dataset includes all key information required for our analysis, such as weekly self-study hours, subject-level scores, and additional relevant variables.
- **C – Current:** The data is collected immediately before the analysis.
- **C – Cited:** The dataset comes directly from the university own data collection.

3. Process

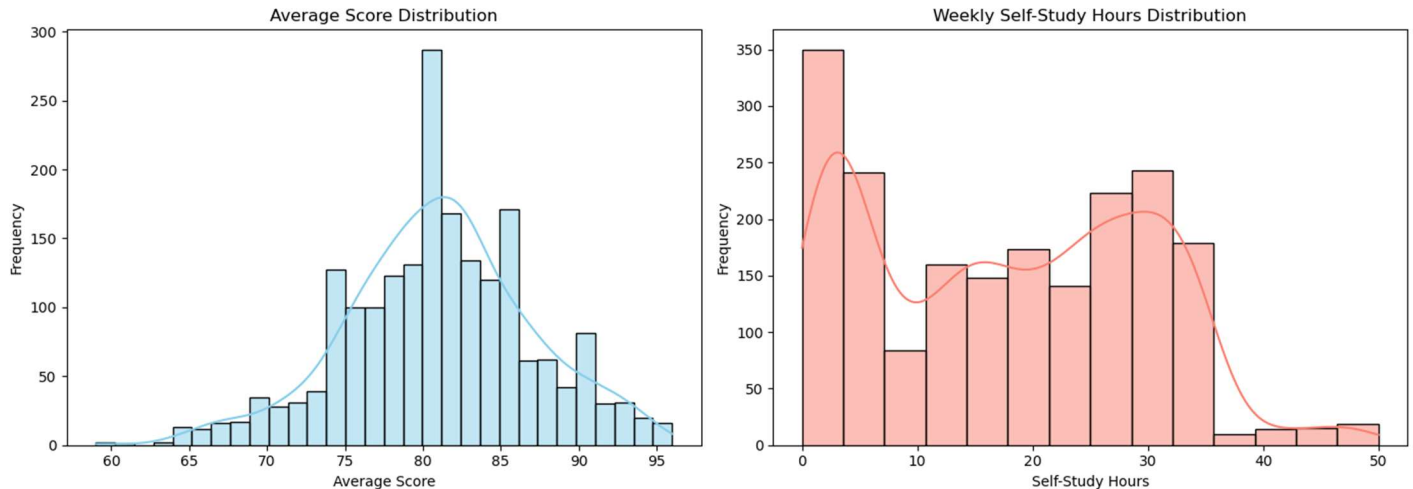
We produced descriptive statistics for the numerical variables and assessed the dataset for missing values. As the data was clean and demonstrated strong integrity, no further cleaning or transformation steps were necessary.

4. Analysis and Share

Exploring the relationship between average score and self-study hours

In this step, we examine whether students who study more tend to achieve higher academic performance. First, we calculate each student's average score across all subjects. We then visualize the distribution of these average scores alongside self-study hours to assess whether they follow a normal or non-normal pattern.

The distributions of the two variables are as follows:



Although it is evident that neither variable follows a normal distribution, we applied the Shapiro–Wilk test to confirm this mathematically:

- Average Score: $W = 0.992$, $p < 0.000000004$
- Self-Study Hours: $W = 0.941$, $p < 0.000000000000000000000002$

Interpretation: both p-values are far below the significance threshold ($\alpha = 0.05$), indicating strong evidence against the null hypothesis of normality. Therefore, Average Score and Weekly Self-Study Hours **are not normally distributed**. This justifies the use of **non-parametric methods like Spearman's rank correlation** for further analysis.

To assess the relationship between students' average academic performance and their weekly self-study hours, we applied Spearman's rank correlational non-parametric test suitable for non-normally distributed data:

- Spearman correlation coefficient (ρ): 0.458
- p-value: 1.772×10^{-104}

Interpretation: the positive coefficient ($\rho = 0.458$) indicates a moderate positive association - students who spend more time on self-study tend to have higher average scores. The extremely small p-value provides strong statistical evidence that this relationship is not due to chance and is significant at any conventional α level. This suggests that self-directed study time is meaningfully linked to academic success in this dataset.

More hypothesis testing...

Although we have already answered the main question - namely, that there is a significant relationship between self-study hours and subject scores - we will continue

exploring the dataset by performing a few additional hypothesis tests (once we've convinced Python to wake up and load the libraries :)).

One-sample t-test on math scores

In this step, we perform a one-sample t-test to determine whether the average math score of students is significantly different from a hypothesized population mean of 83. The output includes:

- T-statistic: measures how far the sample mean is from the hypothesized mean in terms of standard error. This value was 1.528.
- P-value: indicates the probability of observing the data if the null hypothesis is true. A small p-value (typically < 0.05) suggests rejecting the null hypothesis. This value was 0.127, so we can't reject the null hypothesis.

Comparing subject scores between genders using two-sample t-tests:

In this step, we perform independent two-sample t-tests to compare the average scores of male and female students across multiple subjects to test whether there is a statistically significant difference in mean scores between genders for each subject.

The null hypothesis for each test is: H_0 : there is no difference in average [subject] score between male and female students. We use a significance level of $\alpha = 0.05$ to make decisions. The results by subjects:

- Math: reject the null hypothesis
- History: fail to reject the null hypothesis
- Physics: reject the null hypothesis
- Chemistry: fail to reject the null hypothesis
- Biology: fail to reject the null hypothesis
- English: fail to reject the null hypothesis
- Geography: fail to reject the null hypothesis

One-proportion z-test for grade A achievement

We now assess whether the proportion of students achieving Grade A (score ≥ 90) in each subject is significantly greater than 30%.

The null Hypothesis: H_0 : the proportion of students achieving Grade A in [subject] is ≤ 0.3 . Alternative Hypothesis: H_1 : The proportion is > 0.3 .

The results by subjects:

- Math: reject the null hypothesis
- History: fail to reject the null hypothesis
- Physics: reject the null hypothesis
- Chemistry: fail to reject the null hypothesis
- Biology: fail to reject the null hypothesis
- English: fail to reject the null hypothesis
- Geography: fail to reject the null hypothesis

Two-proportion z-test between genders for grade A achievement

We now test whether male students are significantly less likely than female students to achieve Grade A (score ≥ 90) in each subject.

The null Hypothesis: H_0 : the proportion of male students achieving Grade A in [subject] is \geq that of female students. Alternative Hypothesis: H_1 : the proportion of male students is $<$ that of female students.

The results by subjects:

- Math: fail to reject the null hypothesis
- History: fail to reject the null hypothesis
- Physics: fail to reject the null hypothesis
- Chemistry: fail to reject the null hypothesis
- Biology: fail to reject the null hypothesis
- English: reject the null hypothesis
- Geography: fail to reject the null hypothesis

5. Act

Now that we have confirmed a significant relationship between self-study hours and subject scores, and since this work forms part of an advanced statistical methods case study, we can incorporate it directly into the statistics subject matter.