BEYOND THE NUMBERS: UNVEILING THE TRUTH

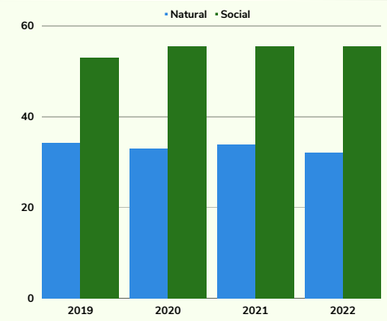
VIETNAMESE NATIONAL HIGH SCHOOL EXAM SCORES

Instructor Tran Duy Hien | Introduction to Data Analysis - Group 1 | Le Thuy Nguyen | Nguyen Thi Thanh Mai | Tran Nguyen Hanh Nguyen

1. MOTIVATION

In recent years, a fascinating trend has emerged: **a higher proportion of candidates are choosing social science (SS) subject combinations compared to natural science (NS) combinations**.

This has sparked our curiosity to explore the relationship between the proportion of candidates choosing subject combinations and its impact on average scores of the National High School Exam among provinces/cities.



The percentage of candidates choosing the Social Science exam has been overwhelming for the past 4 years.

2. RESEARCH QUESTIONS

1. Whether the media press gives us evidence-based information about the distinctive difference in subject composition proportion?
2. Why is there a big difference in students' tendencies between Social Science and Natural Science composition in different provinces/cities?
3. Within each combination of Social Sciences and Natural Sciences subjects, which subject shows the highest correlation with the average scores?

3. METHODOLOGY

Dataset:

- The sample is very large (National High School Exam in Vietnamese over three years in most provinces/cities).
- Collected from Kaggle with reliable sources about data science.

Tests: Confident interval, Density fitting, One sample hypothesis testing, Two-sample hypothesis testing, Correlation between variables.

4. DATA ANALYSIS

Question 1: The reliability of media press

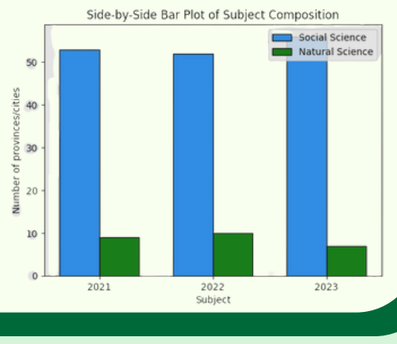
Let Pss be the proportion of students choosing Social Science in each province.

Proportion test

In 2020, 53/62 provinces/cities have
In 2021, 52/63 provinces/cities have
In 2022, 56/63 provinces/cities have

$p_{ss} > 55\%$

- **The proportion of choosing Social Science is always significantly larger than Natural Science and has been consistent over 3 years.**
- **None of the mountainous provinces have a proportion of NS significantly higher than SS.**



Question 2: Testing some our claims explaining why more candidates choose Social Science (SS) than Natural Science (NS)

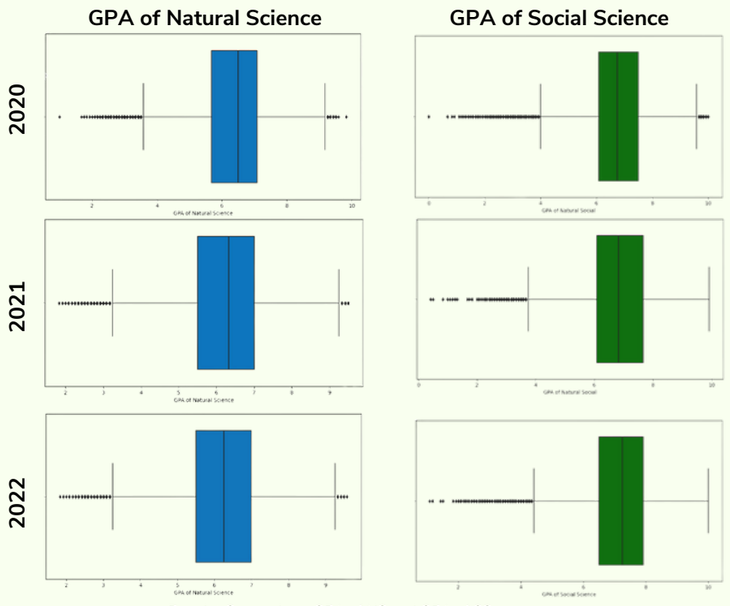
Claim 1: It is easier to get higher score in taking Social Science combination

95% Confidence Interval

	GPA of Natural Science	GPA of Social Science
2020	(6.338, 6.346)	(6.546, 6.552)
2021	(6.247, 6.255)	(6.603, 6.609)
2022	(6.172, 6.180)	(6.904, 6.910)

In each year, the GPA for SS is **consistently higher** than the GPA for NS. This suggests that, on average, **students perform better in SS compared to NS.**

Two population variance test



Boxplot Comparison: GPA of NS and GPA of SS

The consistent **rejection of the null hypothesis** indicates that the population variance in **GPA of NS is consistently greater** than the population variance in **GPA of SS.**

There is evidence to suggest that it is easier to get high score in Social Science.

Claim 2: Choosing Natural Science combination is more prone to paralysis score (<1)

Hypothesis Tests Concerning Two Population Proportions

- Let P1 is the proportion of students getting paralysis score of Natural Science.
- Let P2 is the proportion of students getting paralysis score of Social Science.

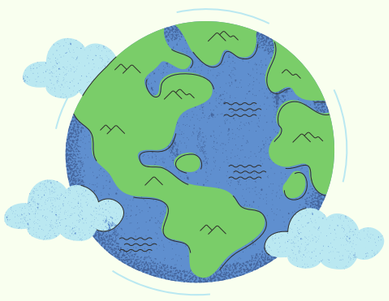
$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 > 0$$
$$TS: Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p_c}(1-\widehat{p_c})(\frac{1}{n_1}+\frac{1}{n_2})}}$$
$$RR: Z \geq Z_\alpha = Z_{0.05} = 1.96\ (\alpha = 0.05)$$

2020: p-value = 0.5968
2021: p-value = 0.8915
2022: p-value = 0.9792

> α = 0.05

→ All three years fail to reject the null hypothesis.

There is no evidence to suggest that students taking Natural Science is prone to get paralysis scores than those taking Social Science.



Claim 3: Mountainous areas students tend to choose SS due to unfavorable studying conditions and worse academic performance

**The list of mountainous provinces was collected from Vietnamnet newspaper.*

Hypothesis Testing Concering Two Population Means

$m_1 -$ *AV Subject (Delta)*
$m_2 -$ *AV Subject (Mountainous)*
$H_0: m_1 - m_2 \leq 0$
$H_a: m_1 - m_2 > 0$
$\alpha = 0.05$
$RR: t > t_\alpha = 1.6449$

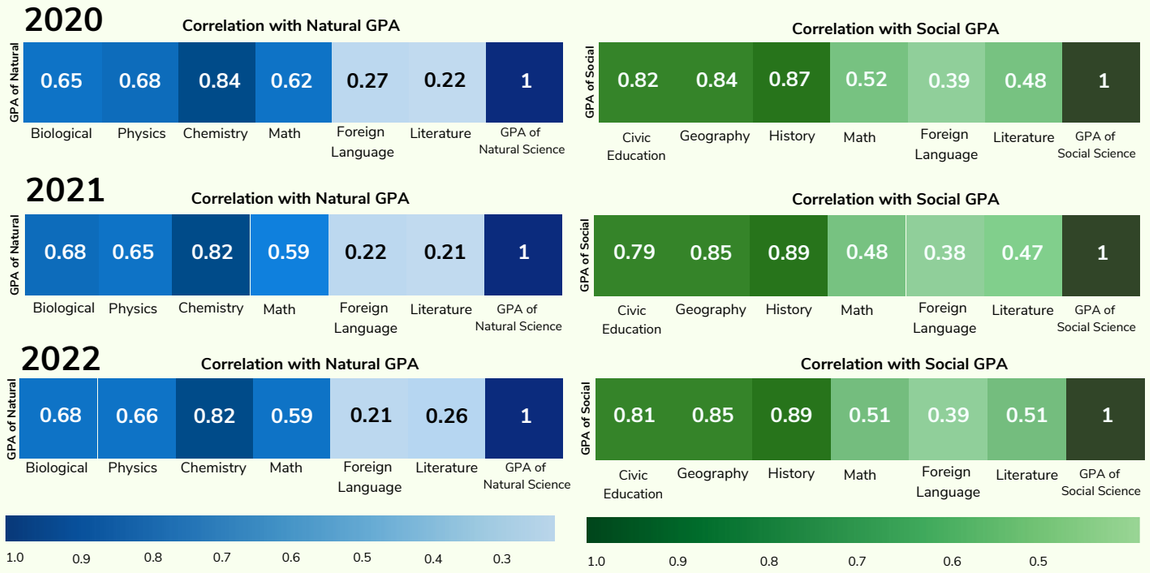
2020: t-score = -3.54
2021: t-score = -3.64

< t = 1.6449

→ Only NS in 2021 and 2022 fail to reject the null hypothesis

- There is weak evidence to suggest that the mean score in NS of delta areas was higher than the mountainous areas.
- Scores of 3 main subjects (Math, Literature & English) have more impact than NS and SS in the mean score and rank of provinces.

Question 3. The correlation between subjects and the average scores



Natural Science

- Chemistry has a closer correlation with the GPA of the NS combination than Biology and Physics.

Social Science

- Civic Education, Geography, and History have quite approximately similar correlations with the GPA of the SS combination.

There is a big difference in the performance of students within Natural Science and Social Science subjects.

CONCLUSION

- The emerging trend to choose SS occurs not only in mountainous areas but also in delta provinces.
- There is not enough evidence to suggest that the proportion of subjects' composition (NS vs SS) has an impact on the average scores and ranks of provinces/cities.
- Scores of 3 main subjects (Math, Literature, & English) have more impact than NS and SS in the mean score and rank of provinces.
- SS witnesses a more strong correlation in each subject composition to overall GPA than NS.

5. DISCUSSION & LIMITATIONS

Discussion

- The **dominant proportion** of Social Science would badly affect **the labor market** of Vietnam (like STEM and technology segment), which beg the question of considering other factors such as region, the quality of education, etc. to evaluate have stimulus program for Natural Science.
- The **different correlations** in each subject's performance of GPAs in Natural Science and Social Science combinations suggest that specific subjects significantly impact students' overall GPAs. This insight can help educators identify areas that may need more attention and support to improve students' overall performance.

Limitations

- **Sample Representativeness:** While the dataset is large, it may not fully represent the entire population of high school students in Vietnam.
- **Causation vs. Correlation:** The statistical tests used can help identify relationships between variables, but they do not establish causation.
- **Timeframe:** The analysis focuses on three years (2020, 2021, and 2022), additional years could provide a more comprehensive understanding of trends and changes in student behavior over time.
- **Subjective Factors:** The decision-making process of students in choosing subject combinations could be influenced by subjective factors such as personal interests, parental guidance, or career aspirations, which may not be fully captured in the data.

References:
https://www.kaggle.com/datasets/nguyenthicamlai/national-high-school-exam-in-vietnamese
https://giaoduc.net.vn/vi-sao-nhieu-thi-sinh-chon-to-hop-khoa-hoc-xa-hoi-hon-khtn-khi-thi-tot-nghiep-post227639.gd
https://meeymap.com/tin-tuc/ban-do-viet-nam