Do MDS students really avoid choosing Advanced Algorithms course?

arbatyrov@edu.hse.ru

August 11, 2023

Abstract

This is an attempt to apply hypothesis testing statistical inference to resolve the following Matvey's concern:

> "Guys, you are kidding, only 5 people chose advanced algorithms... From my very subjective point of view algorithms is a crucial thing, so don't be afraid to choose this course. And, really, what are you afraid of?.. I hope some people will change their minds. I think, some people have some [irrational] fear..."

Disclaimer: this is neither a root cause analysis as to why some students choose to enrol in the Advanced Algorithms course, while some do not, nor is it a motivation analysis behind decision-making.

Discussion

Let's assume that students make their choice unconsciously, or blindly, by tossing a coin: if they obtain a head, they enrol in the course, and if they obtain a tail, they don't. So, each student's choice can be modelled as a Bernoulli process with two outcomes. Now, assuming that students make their choice independently, the sum of these Bernoulli processes is a random variable with Binomial distribution: $X \sim Bin(n, p)$, where $p = 0.5$ (fair coin). Well, some students might be friends or colleagues, or otherwise be not independent, and make identical decisions, but anyway for the sake of simplicity let's assume that a sample $(x_i) \overset{\text{i.i.d.}}{\sim} X$.

Thanks to our dear manager of study office Kristina, we know that there are $n = 64$ students in the Fall2022 cohort, out of which $k = 5$ students have enrolled in the Advanced Algorithms course. Is this number expected or unusual? Well, let's check it with a binomial test for significance level $\alpha = 5 \times 10^{-2}$.

$\mathcal{H}_0 : p = 0.5$ – students do not have any bias in favor of or against Advanced Algorithms course and enrol blindly,

$\mathcal{H}_1 : p < 0.5$ – students have some prior information to be biased *against* Advanced Algorithms course.

To calculate the probability of having $k = 5$ or fewer students enrolled, we need to find the p-value for this k. It can be done manually, recalling that probability mass function for Binomial distribution is given by $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, so p-value $= P(X \leq 5) = CDF(5) = \sum_{k=0}^{5} \binom{64}{k} p^k (1-p)^{64-k}$. But it's much easier to do in Python – see Listing 1.

The obtained p-value $\approx 4.5 \times 10^{-13} \ll \alpha = 5 \times 10^{-2}$, which means that it is extremely unlikely to have only 5 or even fewer students enrolled under $\mathcal{H}_0$, i.e. if students make their choice unconsciously. So, we have to reject the null-hypothesis in favour of the alternative $\mathcal{H}_1$: students have some prior information to be biased *against* Advanced Algorithms course and avoid it. The 95% confidence interval for the number of blindly enrolling students is $[24, 40]$, thus those 5 students who have enrolled are well outside of this CI. The real probability (fraction of enrolled students) is $5/64 = 0.078125 \approx 7.8\%$, which is much lower than the expected $p = 0.5 = 50\%$. So, 5 students is a statistically significant result.

To conclude, it looks like MDS students have some bias *against* the Advanced Algorithms course indeed and consciously avoid enrolling in it. At least, it applies to Fall2022 cohort. Given this result, it is not clear how to convince at least another 19 students to change their mind and "stop being afraid".

```python
import numpy as np
from scipy.stats import binom, binomtest
import matplotlib.pyplot as plt
plt.rcParams['text.usetex'] = True # if LaTex is needed in plots

n = 64
print(f'Number of Fall2022 students: {n = }')
p = 0.5
print(f'Probability of blind enrolment: {p = }')
mu = round(n * p)
print(f'Number of Fall2022 students who might enrolled in Advanced Algorithms course following the blind choice: {mu = }')
alpha = 0.05
print(f'Significance level: {alpha = }')
print()

print("H_0: p = 0.5 -- students do not have any bias in favor of or against Advanced Algorithms course and enrol blindly")
print('H_1: p < 0.5 -- students have some prior information to be biased against Advanced Algorithms course')
k = 5
print(f'Number of enrolled students in fact: {k = }')
H_1 = 'less'
result = binomtest(k=k, n=n, p=p, alternative=H_1)
p_val = result.pvalue
print(f'{p_val = :.1e}')
test_text_start = f'Having only {k} or even fewer students enrolled is'
test_text_end = 'if they make a blind choice indeed.'
print(f'{test_text_start} unusual (reject H_0), {test_text_end}') if p_val < alpha else print(f'{test_text_start} expected (do not reject H_0), {test_text_end}')
print()

# 95% confidence interval for the number of students who might enrol in Advanced Algorithms course
# Binomial distribution of number of enrolled students under H_0
X = binom(n=n, p=p)
k_lo = round(X.ppf(alpha / 2))
k_up = round(X.ppf(1 - alpha / 2))
print(f'95% confidence interval for the number of students, if enrolled unconsciously: [{k_lo}, {k_up}]')
print(f'Fall2022 cohort students who enrolled in fact: {k}')
# As percentage
print(f'95% confidence interval for the percentage of students: [{k_lo / n * 100:.1f}%, {k_up / n * 100:.1f}%]')
print(f'Fall2022 cohort percentage: {k / n * 100:.1f}% ({k}/{n})')
print('This is in CI.') if k_lo <= k <= k_up else print('This is not in CI.')

# Some plotting
x = np.arange(0, n + 1)
x_conf = np.arange(k_lo, k_up + 1) # Confidence interval
plt.figure(figsize=(8, 5))
plt.vlines(x, 0, X.pmf(x), 'C7', ls='--', lw=0.5)
plt.plot(x, X.pmf(x), 'C0o', label='$\mathcal H_0$ (blind choice)')
plt.plot(x_conf, X.pmf(x_conf), 'C3o', label=f'$95\%$ CI for $\mathcal H_0$: $n \in [{k_lo}, {k_up}]$')

# Binomial distribution of number of enrolled students under H_1 for k
Y = binom(n=n, p=k/n)
plt.vlines(x, 0, Y.pmf(x), 'C7', ls='--', lw=0.5)
plt.plot(x, Y.pmf(x), 'C2o', markersize=5, alpha=0.5, label='$\mathcal H_1$ (conscious choice)')
plt.plot(k, 0, 'C1o', label=f'{k} students enrolled in fact')
plt.title('Binomial distributions of number of students enrolled in Advanced Algorithms course')
plt.xlabel('$n$')
plt.ylabel('pmf$(n)$')
plt.legend()
plt.grid(lw=0.25)
plt.tight_layout()
plt.savefig('plot.png', dpi=150)
plt.show();
```

Listing 1: Binomial test Python script

```
Number of Fall2022 students: n = 64
Probability of blind enrolment: p = 0.5
Number of Fall2022 students who might enrolled in Advanced Algorithms course following the blind choice: mu = 32
Significance level: alpha = 0.05

H_0: p = 0.5 -- students do not have any bias in favor of or against Advanced Algorithms course and enrol blindly
H_1: p < 0.5 -- students have some prior information to be biased against Advanced Algorithms course
Number of enrolled students in fact: k = 5
p_val = 4.5e-13
Having only 5 or even fewer students enrolled is unusual (reject H_0), if they make a blind choice indeed.

95% confidence interval for the number of students, if enrolled unconsciously: [24, 40]
Fall2022 cohort students who enrolled in fact: 5
95% confidence interval for the percentage of students: [37.5%, 62.5%]
Fall2022 cohort percentage: 7.8% (5/64)
This is not in CI.
```

Listing 2: Result of running Python script

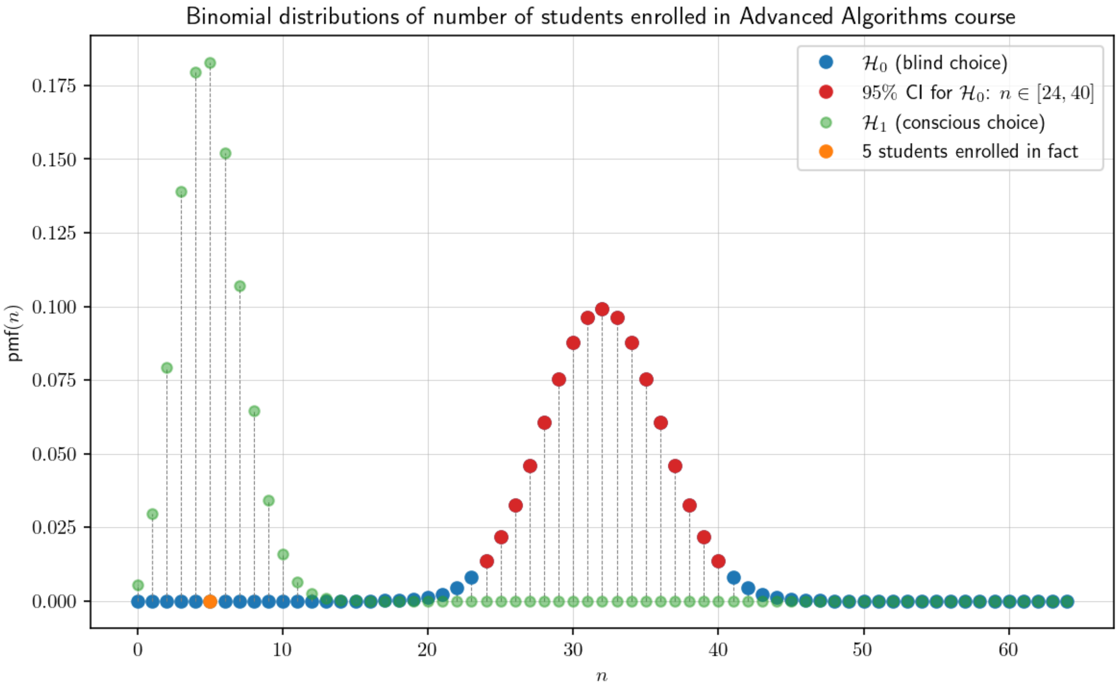

Figure 1: Plots of Binomial distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$