Problem Set 1

Cem Kozanoglu

September 2025

This document describes and explains the results that were obtained from the code in pset1.ipynb. The code was written in Python using the Polars library for data manipulation and analysis, and Matplotlib for visualization.

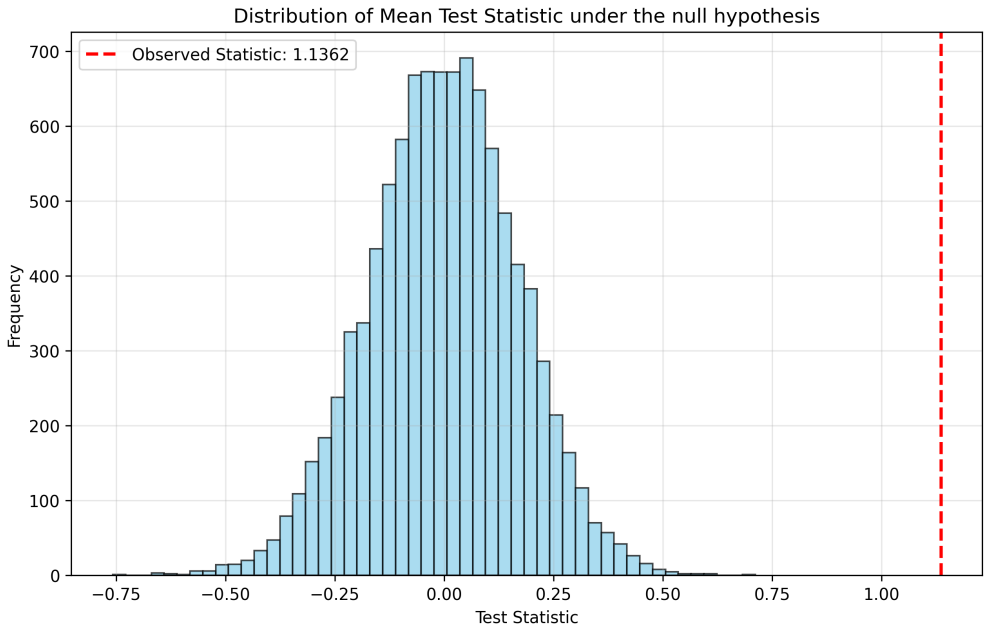
1 Question 1

1.1 Part A

The first test statistic we use is the difference in means between the treatment and control groups, described by T_1 in formula:

$$T_1 = \bar{Y}_T - \bar{Y}_C$$

I generated 10,000 simulations of the test statistic under the null hypothesis of no treatment effect. The histogram below shows the distribution of the simulated test statistics, with a red dashed line indicating the observed test statistic from the actual data.



While the trials converged to a normal distribution around 0 (as we might have expected), our observed T-statistic was 1.1362, extremely unlikely given the null hypothesis. The exact p-value for the observed statistic was 0.00, meaning that there were literally no datapoints

in any of our simulations which were as extreme or more extreme than our observed test statistic. This suggests that the treatment had a statistically significant effect on earnings reported after one year.

1.2 Part B

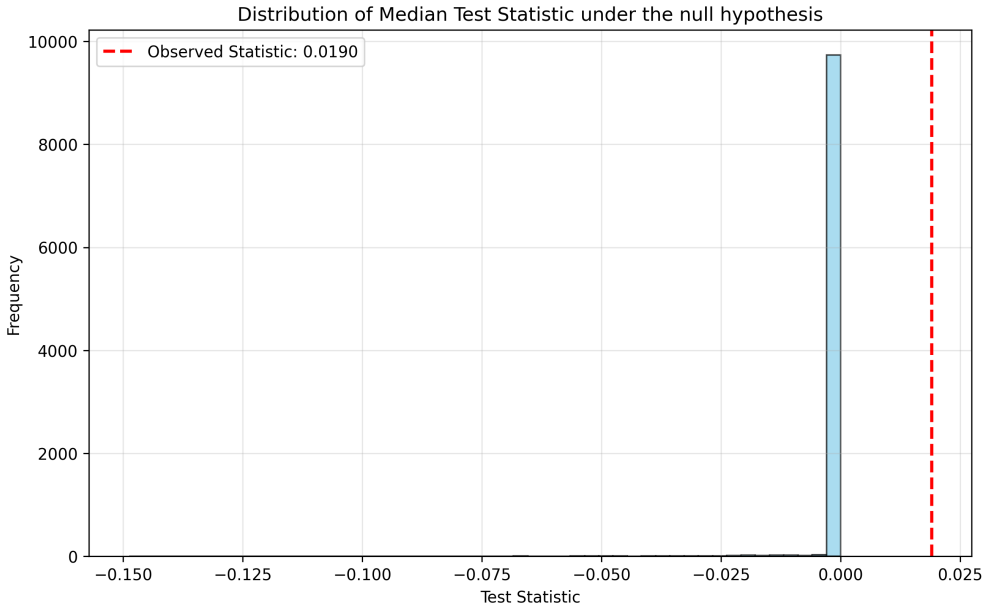
In most cases the difference in means will be the most important statistic to consider when trying to grasp the effects of a treatment. However there may be cases where the mean is not the best measure of central tendency, such as when there are significant outliers in the data, in which case the median may be a more appropriate measure to consider.

I don't think that this example is likely to be one of these cases, given my prior that job training is likely to have fairly uniform effects across the population. However, it certainly won't harm our understanding to compute the difference in medians as a second test statistic.

We thus calculate the second test statistic as the difference in medians between the treatment and control groups, described by T_2 in formula:

$$T_2 = \tilde{Y}_T - \tilde{Y}_C$$

I also calculated this statistic for 10,000 simulations under the null hypothesis of no treatment effect. The histogram below shows the distribution of the simulated test statistics, with a red dashed line indicating the observed test statistic from the actual data.



The trials are clustered tightly around 0 which makes sense given that the median is less sensitive to outliers and our high sample size. Our observed statistic was 0.019, which is again very unlikely under the null hypothesis, and the exact p-value was 0.02, meaning that

only 2% of our simulated test statistics were as extreme or more extreme than our observed statistic. While less extreme than the first statistic, this again suggests that the treatment had a statistically significant effect on earnings reported after one year.

1.3 Part C

To further understand the effects of the treatment, we can also look at the ratio of the variances of the treatment and control groups,

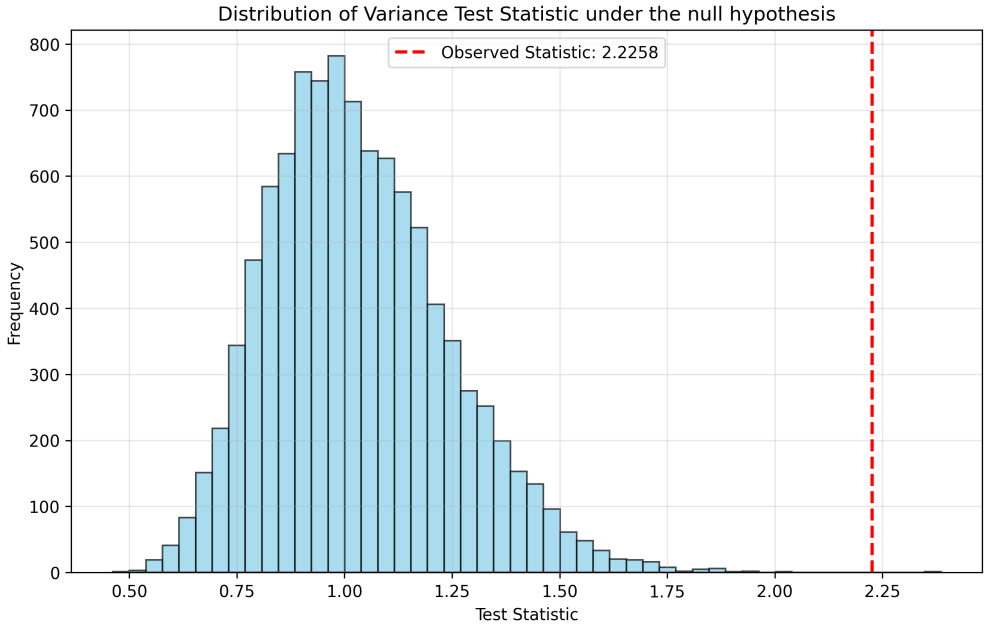
This third test statistic is described by T_3 in formula:

$$T_3 = S_T^2 / S_C^2$$

where:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

I calculated this statistic for 10,000 simulations under the null hypothesis of no treatment effect. The histogram below shows the distribution of the simulated test statistics, with a red dashed line indicating the observed test statistic from the actual data.



In line with our findings for the last two statistics, the trials are clustered tightly around 0. Our observed statistic was 14.97, which is again very unlikely under the null hypothesis, and the exact p-value was 0.03. This again suggests that the treatment had a statistically significant effect on earnings reported after one year, and specifically that it increased the variance of earnings in the treatment group relative to the control group. This could suggest

that while the treatment was beneficial on average, it may have had more varied effects on different individuals, potentially benefiting some significantly while having little to no effect on others.

1.4 Part D

For this section we assign the treatment randomly by using a bernoulli random variable with p = 0.2 to determine wheter a given row will be recorded as a treated or control observation.

The findings from this random assignment are shown in the table below:

Statistic	Combinatoric p-value	Randomized p-value
T_1	0	0
T_2	0.015	0.0143
T_3	0.0001	0.00

As can be seen, our results differ very little from the original combinatoric assignment. This is likely because the high sample size of 5419 is large enough that the random assignment converges to the combinatoric assignment thanks to the Central Limit Theorem.

This type of randomisation would not be ideal in cases with a low sample size, or a low treatment probability, as it would be way more sensitive to variation. For our purposes getting 1000 or 1200 treatment observations does not make a large difference, but if p=0.01 and our number of observations varied from 40 to 60 that would likely have a significant effect on our results.

2 Part 2

Since the treatment and the control groups were assigned randomly, the number of treatment and control variables itself were random, which would not be reflected in our analysis if we chose to randomise our simulations combinatorically by picking 49 observations randomly out of the total 100. It is more reflective of our assignment mechanism if we also randomise our draws for our simulation sample.

Consider that the two methods end up having different probability distributions. For the combinatoric method:

$$P(W) = \binom{100}{49}^{-1} \text{ for } w \text{ s.t. } \sum_{i=1}^{N} \mathbb{1}_{W_i=1} = 49$$

While for the coin-flip method we get:

$$P(W = w) = 2^{-N}, \forall w \in \{0,1\}^N$$

If our sample was significantly smaller or our treatment probability was significantly lower, this difference in distribution would likely have a significant effect on our results, and we may prefer to randomise combinatorically.