

A/B Test and Hypothesis Testing

by Yuditya Artha

What is A/B Test?

Definition

A/B testing is a way to compare two versions of something to find out which version performs better. For example, a data professional might use A/B testing to compare two versions of a web page or two versions of an online ad. A/B testing utilizes statistical methods such as sampling and hypothesis testing.

(Google Advanced Data Analytics, Course 4 - The Power of Statistics)

The Dataset

I use this dataset from Kaggle and here is the link:
https://www.kaggle.com/datasets/sergylog/ab-test-data?select=AB_Test_Results.csv

Tools I used & workflow:

Raw dataset (.csv) > Anaconda > Jupyter Notebook > Python with the following libraries : (Pandas, Numpy, Matplotlib, Seaborn, Stats, proportions_ztest, mannwhitneyu)

I want to answer these questions

Questions

- 1) Do revenue from control and variant groups have different average revenue?
- 2) If the answer of Q1 is yes, is there a difference in revenue performance between control and variant groups??

Frame the Null and Alternative Hypotheses

H_0 (null): There is no difference in revenue between control and variant.

H_1 (alt): There is a difference in revenue between control and variant.

Data Wrangling

What did I do here?

- Import necessary libraries
- Load the dataset
- Have a glimpse of the first 5 rows of the data
- Checking the data size

The data has total of 10.000 rows and 3 columns

1. Imports and Data Loading

Import packages and libraries needed to compute descriptive statistics and conduct a hypothesis test.

```
# Import packages for data manipulation
import pandas as pd
import numpy as np

# Import packages for data visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Import packages for statistical analysis/hypothesis testing
from scipy import stats
```

```
# Load dataset into dataframe
data = pd.read_csv("AB_Test_Results.csv")
```

```
data.head()
```

	USER_ID	VARIANT_NAME	REVENUE
0	737	variant	0.0
1	2423	control	0.0
2	9411	control	0.0
3	7311	control	0.0
4	6174	variant	0.0

```
print("The data size is : " + str(data.size))
print("With total of columns are " + str(data.shape[1]) + " columns and total of rows are " + str(data.shape[0]) + " rows.")
```

The data size is : 30000
With total of columns are 3 columns and total of rows are 10000 rows.

Exploratory Data Analysis (EDA) - Checking nulls and skewness

2. Checking Nulls

```
## A glimpse of descriptive analytics in the dataset
data.describe()
```

	USER_ID	REVENUE
count	10000.000000	10000.000000
mean	4981.080200	0.099447
std	2890.590115	2.318529
min	2.000000	0.000000
25%	2468.750000	0.000000
50%	4962.000000	0.000000
75%	7511.500000	0.000000
max	10000.000000	196.010000

From the result above we see there is no null value from all of the three columns (all the total count is 10000 records)

But let's do another double check, if there are nulls then there should be total record of nulls on the result of this query

```
data.isna().sum()
```

```
USER_ID      0
VARIANT_NAME  0
REVENUE       0
dtype: int64
```

All set, no nulls value exist within the dataset.

What did I do here?

- Have a glimpse of descriptive analytics the dataset with `.describe()`
The result shows no null (all 10000 records are complete and no nulls)
- Confirm the nulls once again with `.isna().sum()`
The data has total of 10.000 rows and 3 columns

3a. Measure the data skewness

```
# Calculate the revenue skewness
print("Skewness:", "{:1.2f}".format(data["REVENUE"].skew()))
from scipy.stats import shapiro

stat, p = shapiro(data["REVENUE"])
print(f"Shapiro-Wilk Test p-value: {p}")
```

Skewness: 64.98

Shapiro-Wilk Test p-value: 8.963551861875658e-115

What did I do here?

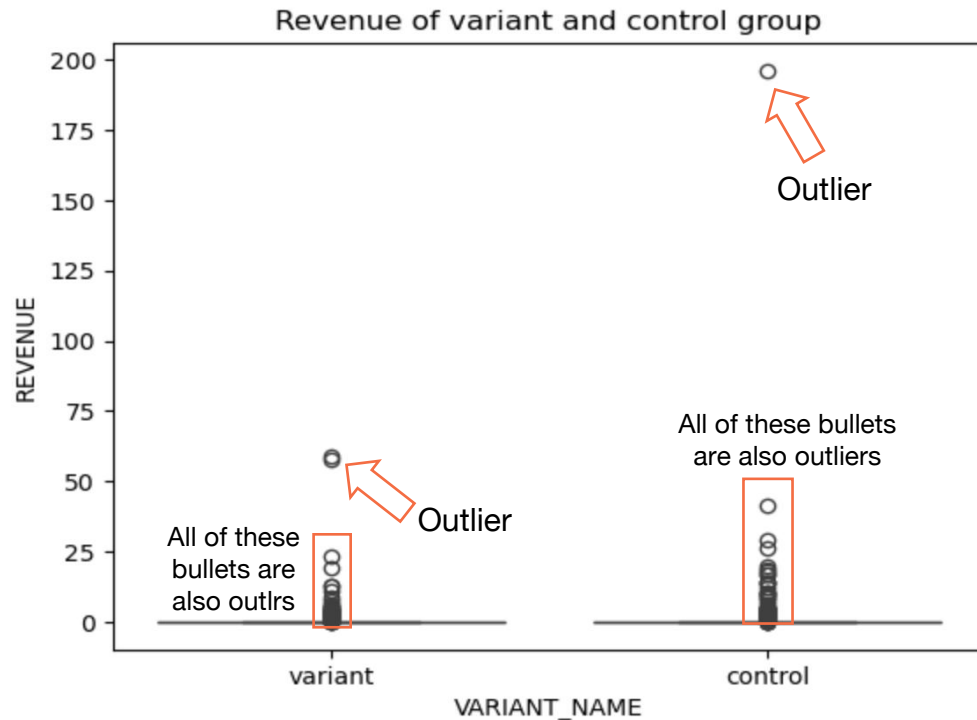
Skewness measures how asymmetric a distribution is. A perfectly symmetric distribution has skewness ≈ 0 . Rules of thumb (skewness) categorization:

- $(-0.5, 0.5)$ —> low or approximately symmetric
- $(-1, -0.5)$ or $(0.5, 1)$ —> moderately skewed.
- Beyond -1 and 1 —> Highly skewed.

For this dataset, the revenue is extremely positive skewed (right-tail skew) because the **skewness is 64,98**

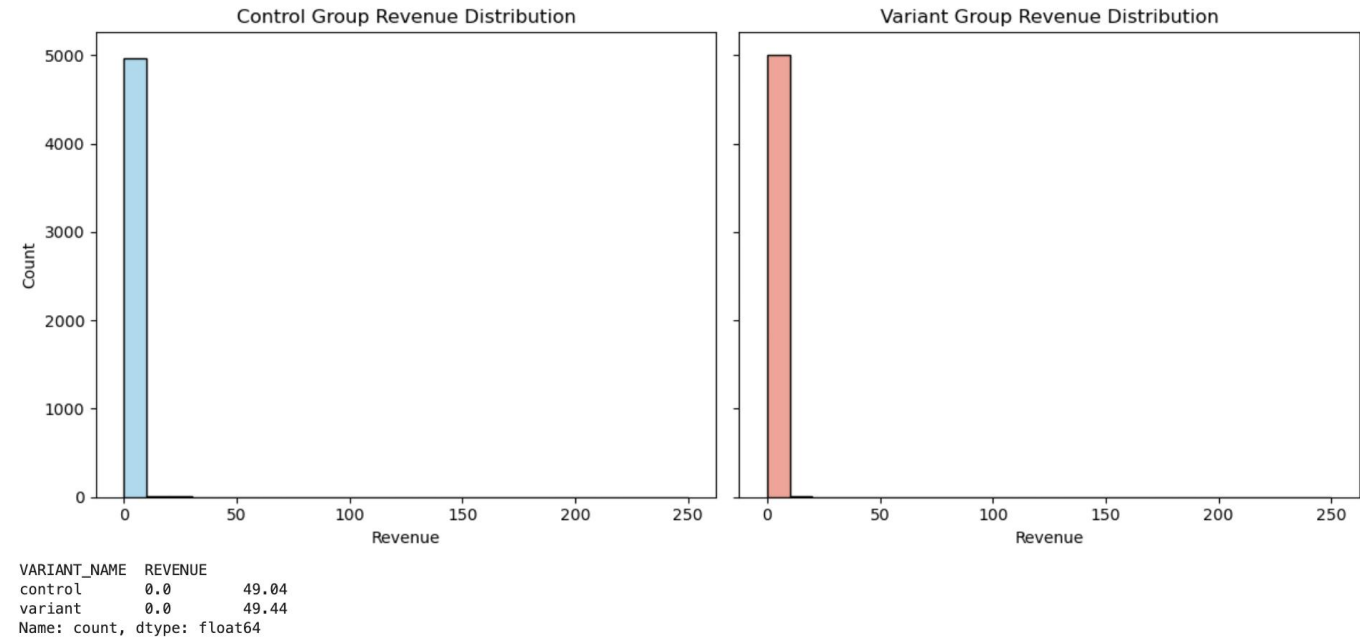
Exploratory Data Analysis (EDA) - Visualize data distribution

```
# Create the boxplot of the data
plt.title('Revenue of variant and control group')
sns.boxplot(data=data, x="VARIANT_NAME", y="REVENUE")
```



The boxplot above is showing the distribution of revenue of both groups (variant and control).

Outliers are dominant and unavoidable, both control and variant have high-value users. The box (IQR) is basically lying on 0.



The entire distribution is jammed near zero, and the y-axis shoots up approximately to 5,000 record counts for both groups.

The histogram plot above is showing that entire distribution is jammed near zero, and the y-axis shoots up to almost 5,000 counts per group.

Insights from EDA process:

1. Almost all users (98,5% of records) generate zero revenue and 0,2% users generate near-zero revenue
2. Almost ~49% users of total users from each group generate zero revenue
3. Although the revenue distribution contains extreme outliers, these likely represent high-value customers

Since the goal is to test real business impact, I retain these values for initial hypothesis testing

Checking Nulls and comparing the average revenue of both groups

4. How different are zero vs non-zero revenue users?

```
ovr_avg_median = data.groupby("VARIANT_NAME")["REVENUE"].agg(['mean','median']) # this include all the zero revenue
paid_avg_median = data[data["REVENUE"]>0].groupby("VARIANT_NAME")["REVENUE"].agg(['mean','median']) # this include only non zero revenue

summary = pd.merge(ovr_avg_median, paid_avg_median, on = "VARIANT_NAME")
summary.columns = ["avg_ovr_revenue", "median_ovr_revenue", "avg_paid_revenue", "median_paid_revenue"]
summary
```

	avg_ovr_revenue	median_ovr_revenue	avg_paid_revenue	median_paid_revenue
VARIANT_NAME				
control	0.1290	0.0	8.0375	2.96
variant	0.0701	0.0	4.8815	2.17

```
print("Result Summary:") ●●●

Result Summary:
The average overall revenue difference of both group is: 0.06
The overall paid revenue difference of both group is: 3.16
The median paid revenue difference of both group is: 0.79
The median of overall revenue are equally 0

#Summary of ARPU and AOV
arpu = data.groupby('VARIANT_NAME')['REVENUE'].mean() # ARPU = average revenue per user (includes zeros)

aov = data[data['REVENUE']>0].groupby('VARIANT_NAME')['REVENUE'].mean() # AOV = average order value (only payers)
print("ARPU:\n", "\n", arpu, "\n=====")
print("AOV (payers):\n", "\n", aov)
```

ARPU:

```
VARIANT_NAME
control    0.1290
variant    0.0701
Name: REVENUE, dtype: float64
=====
```

AOV (payers):

```
VARIANT_NAME
control    8.0375
variant    4.8815
Name: REVENUE, dtype: float64
```

data.describe()		
	USER_ID	REVENUE
count	10000.000000	10000.000000
mean	4981.080200	0.099447
std	2890.590115	2.318529
min	2.000000	0.000000
25%	2468.750000	0.000000
50%	4962.000000	0.000000
75%	7511.500000	0.000000
max	10000.000000	196.010000

Question 1:
Do revenue from control and variant groups have different average revenue?

The answer : **YES!**

- The average revenue of control group is 0,1290
- The average revenue of variant group is 0,0701
- The difference of average value from both group is 0,06

However.....

From the .describe() result, it seems the revenue data is highly skewed and has outliers, which impact the average value of both group. I arrive on this conclusion because:

1. The average revenue value is 0,099
2. But, the revenue median is 0
3. The 75% percentile on revenue column is 0
4. But the max_value of revenue column is 196,01
5. These averages include users who made no purchases (revenue = 0).

The revenue column has a lot of 0 values, which heavily skews the results.

Determining statistical test to use

Before selecting a statistical test.....

Researcher has to simply answer the following six questions, which will lead to correct choice of test.

1. How many independent variables (IV) covary (vary in the same time period) with the dependent variable? *One variable IV, the variant name*
2. At what level of measurement is the independent variable (IV)? *The IV is variant name, which consist two category (or binary)*
3. What is the level of measurement of the dependent variable (DV)? *The DV is revenue value which continuous*
4. Are the observations independent or dependent? *Control and variant group are Independent to each other*
5. Do the comparisons involve populations to populations, a sample to a population, or are two or more samples compared?
A/B Test is comparing two sample scenario (control and variant).
6. Is the hypothesis being tested comparative or relationship? *Comparative, we compare of control or variant group performs differently*

Unpaired 2 sample t - test mean checked all the requirements above.

However, since the revenue data distribution is highly positive skewed, **proportion Z test** or **Mann-Whitney U test** is much more suitable to use.

Reference: Parab, S., & Bhalerao, S. (2010). Choosing the correct statistical test: A decision-making flowchart. International Journal of Hygiene and Environmental Health, [volume(issue)], pages.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996580/> PMID:21170214

 <https://linkedin.com/in/yuditya-artha>

Hypothesis Testing - proportion_ztest

Since the answer of Question 1 is yes, **I proceed to Question 2:**

Is there a relationship between control and variant groups on their conversion rate performance?

The answer is:

failed to reject the null hypothesis, the observed difference is not statistically significant.

From this data I can't say the variant group performed better or worse than control group and the difference of their conversion rate occur by chance.

```
from statsmodels.stats.proportion import proportions_ztest
```

```
conversions = conversion_rate["Conversions"]
totals = conversion_rate["Total"]
```

```
# Run the test
```

```
z_stat, p_value = proportions_ztest(count=conversions, nobs=totals, alternative='two-sided')
print(f"Z-stat = {z_stat:.4f}, p-value = {p_value:.4f}")
```

```
Z-stat = 0.6936, p-value = 0.4879
```

```
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant difference of conversion rate between group and does not occur by chance.")
else:
    print("Fail to reject the null: No significant difference detected of conversion rate between group and the difference occur by chance.")
```

```
Fail to reject the null: No significant difference detected of conversion rate between group and the difference occur by chance.
```

Follow Up Question:

However, is it confirmed that the difference between group is not worth acting on based on the effect size?

For proportion Z-Test, I use this test to answer above question:

1. Absolute difference ($p_2 - p_1$) → in percentage points.
2. Relative difference (relative lift) = $(p_2 - p_1) / p_1$.
3. Cohen's h → standardized measure for comparing two proportions.

Personal Note:

Initially, as I mentioned on previous slide, I considered using a t-test to **compare the average revenue** between the control and variant groups. However, I discovered that the revenue distribution is heavily skewed due to a high number of zero values (i.e., users who did not convert). This violates the normality assumption required for a valid t-test.

To address this, I shifted focus to **comparing the conversion rates** (i.e., the proportion of users who made a purchase), by creating a binary variable paid (1 if revenue > 0, else 0).

Since:

- The variable paid is Bernoulli-distributed (binary),
- The sample sizes are large and roughly balanced (~5,000 records per group),
- And the success/failure counts are sufficient ($np > 5$ and $n(1-p) > 5$),

...I used a two-tailed Z-test for proportions, which is valid under the Central Limit Theorem (CLT).

Hypothesis Testing - proportion_ztest

```
print("=== Conversion (control vs variant) ===") •••

=== Conversion (control vs variant) ===
N_control = 4984 N_variant = 5016
Conversions control = 80 variant = 72
control p1 = 0.016051 (1.6051%)
variant p2 = 0.014354 (1.4354%)
absolute difference (variant - control) = -0.001697 (-0.1697 percentage points)
relative lift = -10.57%
Cohen's h = -0.013877 ~ -0.014
```

```
# Output
print(f"Baseline rate: {p1*100:.2f}%")
print(f"MDE (with current N, 80% power, α=0.05): {mde_pp:.2f} percentage points")
print(f"Required Cohen's h: {effect_size_needed:.4f}")
print(f"Observed Cohen's h: {h_observed:.4f}")
print(f"Required number sample: {required_n:.0f}")
```

```
Baseline rate: 1.61%
MDE (with current N, 80% power, α=0.05): 0.78 percentage points
Required Cohen's h: 0.0560
Observed Cohen's h: 0.0139
Required number sample: 81518
```

Recommendation:

If this were a real business experiment, there are several choices can be called:

1. Increase N massively → run the test longer, get ~81k users per group.
2. Accept lower power → risk missing the effect (more false negatives).
3. Stop testing → if the effect is too small to matter financially, it's not worth chasing.

Full code snippet can be viewed on full notebook that available on my [Github](#)

RESULT:

The Cohen's h (-0,014), which measures the difference between two proportions on a standardized scale is so small, almost negligible.

Meaning the observed lift between control and variant conversion rates is minuscule, even if the drop were real, it's too small to matter for most decisions.

Since both Cohen's h and observed cohen's h (0.2258) < required cohen's (0,4551). The sample size is underpowered for this observed difference with 80% power.

Required_n = 81,518 (16x more data) per group

This is how many samples needed in each group to reliably detect an effect of that small size with 5% significance level & power of 0.8 (80% chance of detecting it if it's real)

In other words, I currently have ~5,000 per group which would need over 16x more data to have a decent chance of finding that tiny effect significant.

The actual absolute difference in conversion rate between control (1.61%) and variant (1.43%) = 0.18 percentage points. Much smaller than MDE (0,4551), so it's no surprise result of p-value wasn't significant (> 5%) & CI straddled zero

I'm still being skeptical and explorative, I perform another test

The Mann-Whitney U Test

Hypothesis Testing - using The Mann-Whitney U Test

B) A/B test using Mann Whitney U Test

```
from scipy.stats import mannwhitneyu

paid_users = data[data['paid'] == True]

control_revenue = paid_users[paid_users['VARIANT_NAME'] == 'control']['REVENUE']
variant_revenue = paid_users[paid_users['VARIANT_NAME'] == 'variant']['REVENUE']

revenue_summary = paid_users.groupby('VARIANT_NAME')['REVENUE'].agg(['count', 'mean', 'median']).reset_index()
revenue_summary.columns = ['Group', 'Count', 'Mean_Revenue', 'Median_Revenue']
print(revenue_summary)
```

	Group	Count	Mean_Revenue	Median_Revenue
0	control	80	8.037500	2.96
1	variant	72	4.881528	2.17

```
stat, p = mannwhitneyu(control_revenue, variant_revenue, alternative='two-sided')

print("Mann-Whitney U statistic:", stat)
print("P-value:", p)
```

Mann-Whitney U statistic: 3356.0
P-value: 0.07924299810603058

```
alpha = 0.05
if p < alpha:
    print("Reject the null hypothesis: There is a significant difference and does not occur by chance.")
else:
    print("Fail to reject the null: No significant difference detected and the difference occur by chance.")
```

Fail to reject the null: No significant difference detected and the difference occur by chance.

Follow Up Question:

However, is it confirmed that the difference between group is not worth acting on based on the effect size? For Mann-Whitney, I use **Rank-biserial effect size** to answer above question

Mann-Whitney U test, also known as the Wilcoxon rank-sum test is:

- A non-parametric alternative to the t-test,
- Compares the ranked values rather than raw values,
- Does not assume normality and is robust to outliers and skewed distributions.

This test evaluates whether the distribution of revenue values differs significantly between the two groups.

Using Mann-Whitney U Test, the answer is **Failed to reject the null hypothesis, there's no strong statistical evidence to conclude that revenue between paying users control and variant group differs significantly.**

I can't say the distribution of revenue values differs significantly between the two groups. The result does not imply that variant group performed better or worse than control group and their difference on revenue just occur by chance.

Hypothesis Testing - The Mann-Whitney U Test

However, is it confirmed that the difference between group is not worth acting on based on the effect size?
For Mann-Whitney, I use Rank-biserial effect size to answer above question:

```
# Mann-Whitney rank-biserial from U (use the U you got for control)
n1 = len(control_revenue)
n2 = len(variant_revenue)
U = stat
# rank-biserial:
rank_biserial = 1 - (2*U)/(n1*n2)

print("\n=== Revenue among paying users (Mann-Whitney effect) ===")
print("Mann-Whitney: paying users control =", n1, "& variant =", n2)
print("Observed U =", U, "p =", mw_p)
print(f"U (for control) = {U}")
print(f"rank-biserial = {rank_biserial:.6f}, negative means effect size of variant > control (if passed control first to mannwhitneyu syntax)")

=== Revenue among paying users (Mann-Whitney effect) ===
Mann-Whitney: paying users control = 80 & variant = 72
Observed U = 3356.0 p = 0.07924299810603058
U (for control) = 3356.0
rank-biserial = -0.165278, negative means effect size of variant > control (if passed control first to mannwhitneyu syntax)
```

Mann-Whitney's U tells whether ranks differ, but I also want a standardized effect size: using **rank-biserial**

Interpretable measure. It ranges roughly from -1 to +1:

- Positive → group 1 tends to have higher values
- Negative → group 2 tends to have higher values

Magnitude interpretation is context-dependent:
~0.1 small, 0.3 medium, 0.5 large (very rough).

Result Interpretation

Negative value (-0.165) means the variant tended to have higher ranks (higher revenue) than control.

This means although fewer people in variant converted (conversion metric), among those who did convert variant payers tended to spend slightly more and but the difference magnitude is modest.

Quick Summary Table of The Project Results

Metric	Control	Variant	Result (p-value)	Significance (p-value > 0,05)
Conversion Rate	1.61%	1.43%	0.488	No
Median Revenue (paying users)	2,96	2,17	0.079	No

Conclusion:

There is no any supporting evidence that using variant enhancement or improvement generating more revenue since both Proportion Z-Test and Mann-Whitney U Test fail to reject the null hypotheses that states there is no difference in revenue between control and variant. From effect size test result, both test tell there is no enough magnitude difference to call action that worth the notice. Bootstrapping the median revenue CI for both group also support that there is inconclusive whether the variant group revenue differs significantly.

Without statistical or business impact, there's no justification for further investment in this variant.






Thank you!

I hope you find this useful! 😊

Lihat data lengkap, kode dan visualisasi di GitHub:

<https://github.com/yudityaarth/A-B-Test-Project>

Questions or Feedback? Please email ke yuditya_artha@outlook.com or DM me 😊

 <https://linkedin.com/in/yuditya-artha>
 <https://github.com/yudityaarth>
 Tableau Public: <https://public.tableau.com/app/profile/yuditya.artha/vizzes>
 +62 818-1865-3000
 yuditya_artha@outlook.com or yuditya.artha@gmail.com