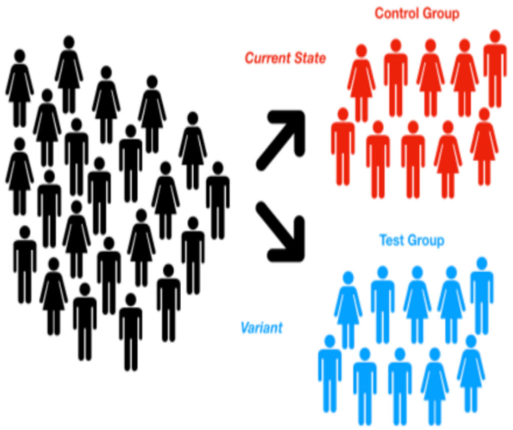
AB Testing for e-commerce company

# The code for calculations and visualizations: https://github.com/annatsf/ab-testing/blob/main/ab-testing.ipynb

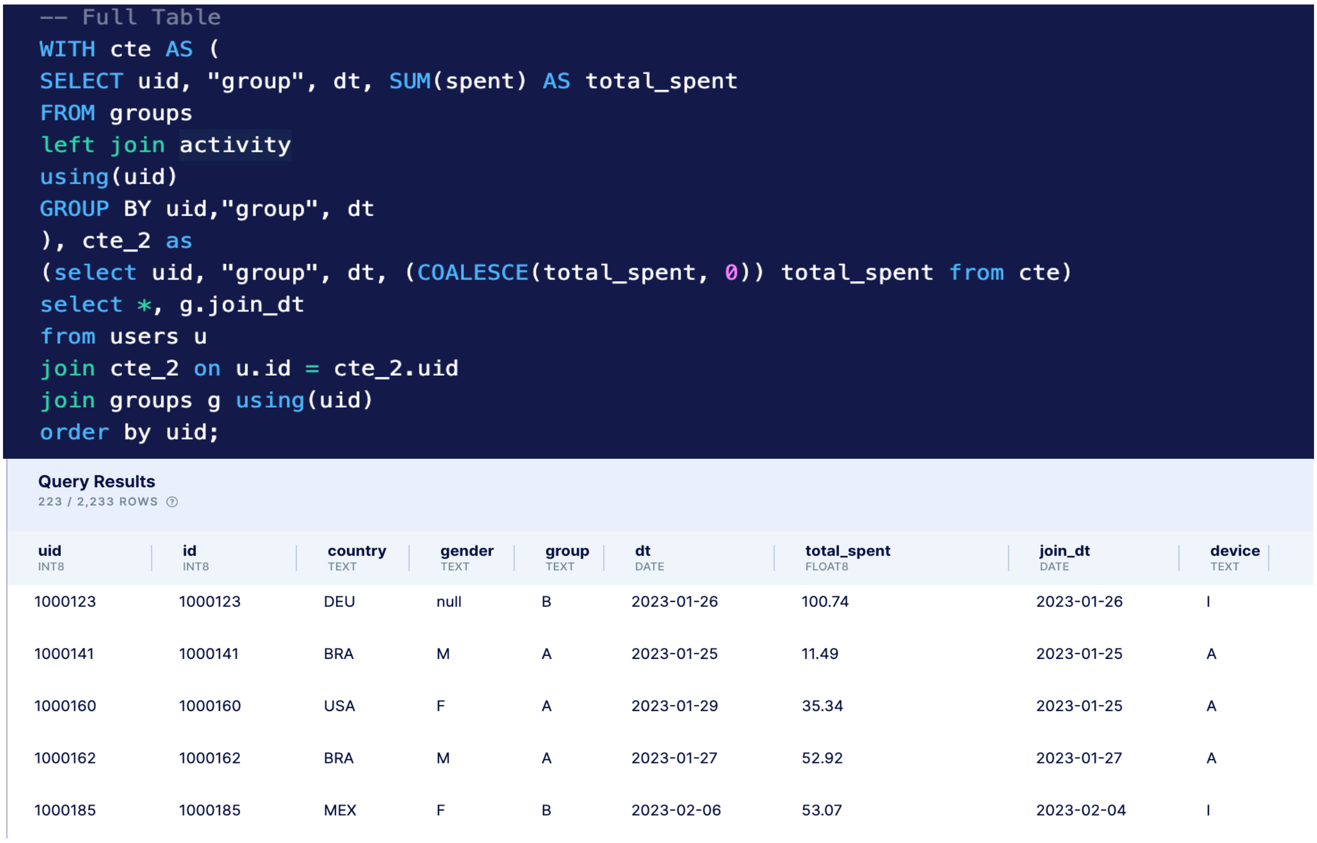
Assume that I am a Data Analyst for an e-commerce company called GloBox. GloBox is an online marketplace that specializes in sourcing unique and high-quality products from around the world.
We believe that shopping should be an adventure, and we want to bring the world to your doorstep. From exotic spices and rare teas to handmade jewelry and textiles, we have a curated selection of products that you won't find anywhere else.

The Growth team decides to run an A/B test that highlights key products in the food and drink category as a banner at the top of the website. The control group does not see the banner, and the test group sees it.

Control and treatment (test) groups



Our data set:



```
-- Full Table
WITH cte AS (
SELECT uid, "group", dt, SUM(spent) AS total_spent
FROM groups
left join activity
using(uid)
GROUP BY uid,"group", dt
), cte_2 as
(select uid, "group", dt, (COALESCE(total_spent, 0)) total_spent from cte)
select *, g.join_dt
from users u
join cte_2 on u.id = cte_2.uid
join groups g using(uid)
order by uid;
```

Query Results
223 / 2,233 ROWS

uid	id	country	gender	group	dt	total_spent	join_dt	device
INT8	INT8	TEXT	TEXT	TEXT	DATE	FLOAT8	DATE	TEXT
1000123	1000123	DEU	null	B	2023-01-26	100.74	2023-01-26	I
1000141	1000141	BRA	M	A	2023-01-25	11.49	2023-01-25	A
1000160	1000160	USA	F	A	2023-01-29	35.34	2023-01-25	A
1000162	1000162	BRA	M	A	2023-01-27	52.92	2023-01-27	A
1000185	1000185	MEX	F	B	2023-02-06	53.07	2023-02-04	I

The setup of the A/B test is as follows:

1. A user visits the GloBox main page and is randomly assigned to either the control or test group. This is the join date for the user.
2. The page loads the banner if the user is assigned to the test group, and does not load the banner if the user is assigned to the control group.
3. The user subsequently may or may not purchase products from the website. It could be on the same day they join the experiment, or days later. If they do make one or more purchases, this is considered a "conversion".

Key performance indicators (KPIs)

A/B Tests: Measure impact of changes on KPIs

- KPIs — metrics are important to an organization:
  - Primary goal: increase revenue
  - Better metric: conversion rate.

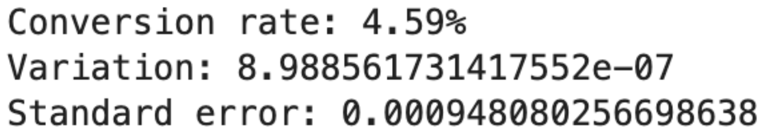
Conversion rate important:

- More granular than overall revenue
- Strong measure of growth
- Directly related to the test
- Potential early warning sign of problems

    - Sensitive to changes in the overall ecosystem

Data variability.

The conversion rate doesn’t vary a lot among users - it will be easier to detect a change



```
Conversion rate: 4.59%
Variation: 8.988561731417552e-07
Standard error: 0.000948080256698638
```

Formulating a hypothesis.

First things first, we want to make sure we formulate a hypothesis. This will make sure our interpretation of the results is correct as well as rigorous.

Given we don't know if the new design will perform better or worse (or the same?) as our current design, we'll choose a two-tailed test

H0:p=p0
Ha:p≠p0

where p and p0 stand for the conversion rate of the and old design, respectively. We'll also set a confidence level of 95%:

α=0.05
The α value is a threshold we set, by which we say "if the probability of observing a result as extreme or more (p-value) is lower than α, then we reject the null hypothesis". Since our α=0.05 (indicating 5% probability), our confidence (1 - α) is 95%.

- Hypothesis that control & treatment have the same impact on the response
    - loaded the banner does not improve conversion rate
    - Any observed difference is due to randomness

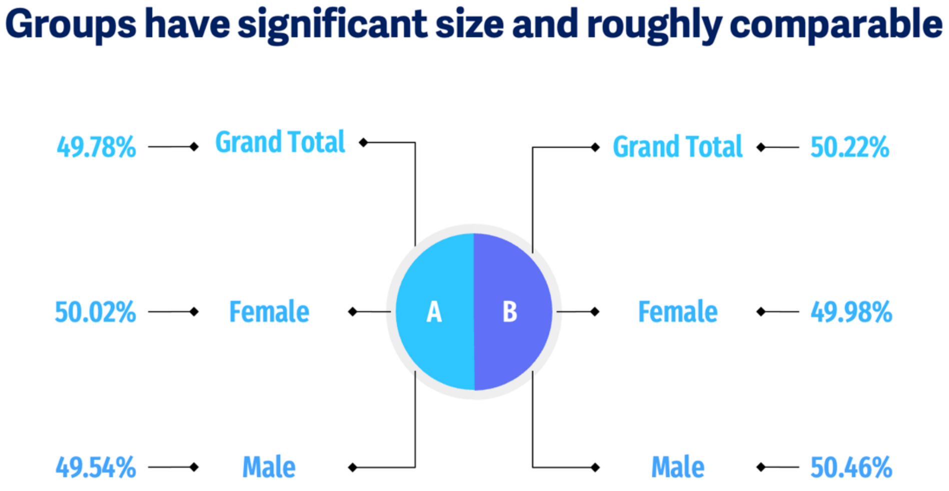
- Rejecting the Null Hypothesis
    - Determine their is a difference between the treatment and control group
    - Statistically significant result

Analyzing the A/B test results

Confirming our test results

Our target randomization unit of the test – individual users. It makes a convenient experimental unit and compare the average spent and conversion rate across two groups.
Our groups have significant size and roughly comparable and have similar demographics:

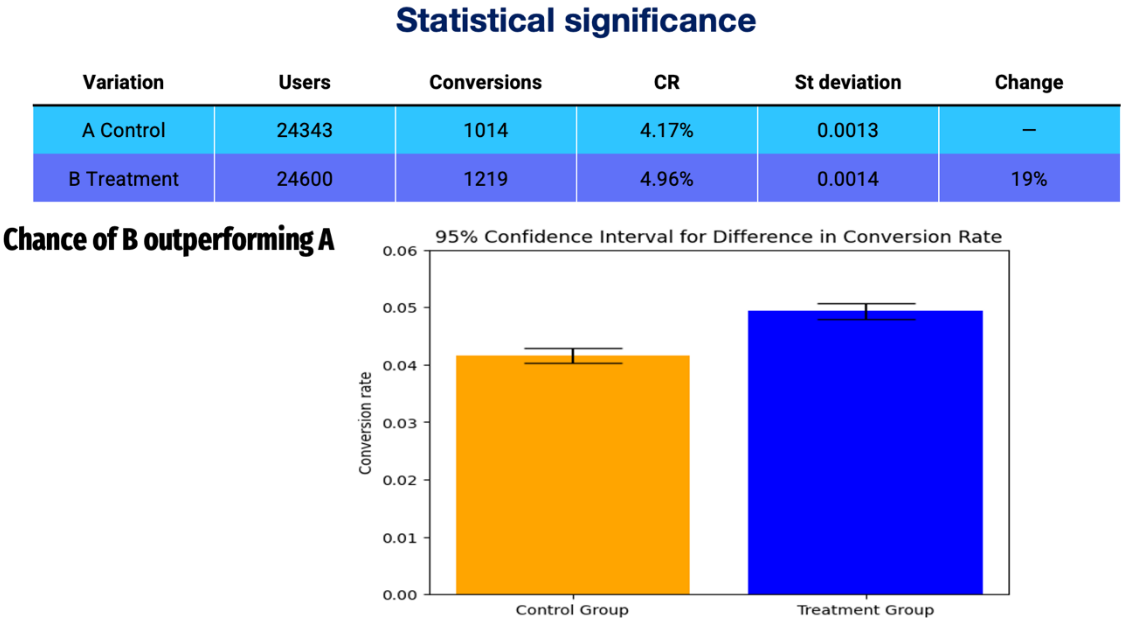
	Female		Male	
Groups	Android	IOS	Android	IOS
A (control)	49,80%	50,42%	49,75%	49,19%
B (treatment)	50,20%	49,58%	50,25%	50,81%



Groups have significant size and roughly comparable

Checking for statistical significance

Here results of the calculation of the size of the test and control groups and calculate their respective conversion rates and standard deviations (represents how ‘spread-out’ the data points are).



Statistical significance

Variation	Users	Conversions	CR	St deviation	Change
A Control	24343	1014	4.17%	0.0013	–
B Treatment	24600	1219	4.96%	0.0014	19%

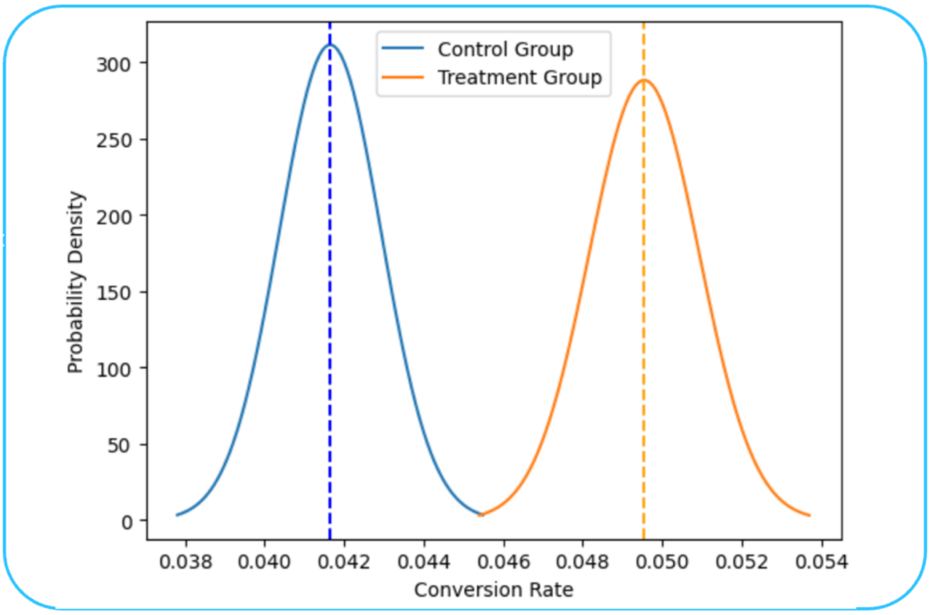
Chance of B outperforming A

Judging by the stats, it does look like our two designs performed very similarly, with our new design performing slightly better, approx. 4.17% vs. 4.96% conversion rate.

The conversion rates for our groups are indeed very close. Also note that the conversion rate of the control group is lower than what we would have expected given what we knew about our avg. conversion rate (4.59%) This goes to show that there is some variation in results when sampling from a population.
The treatment group has shown a higher level of impact than the control. This means the treatment has performed by 19% than the control group. In other words, the alternative hypothesis is more impactful than the null hypothesis. And based on this, the feature should be rolled out into full production.

Conversion rate distributions.
Here, we will visualize the test and control conversion rates as distributions. Additionally, viewing the data in this way can give a sense of the variability inherent in our estimation.

Distribution of Conversion Rates



We have two normal distributions for the mean conversion rate of control and treatment groups. The dashed lines represent the mean conversion rate for each group, and the distance between the two
lines represents the mean of the difference between the control and treatment group. In probability theory, the sum of the normally distributed independent random variables is also normally distributed.

After running our experiment, we get a resulting conversion rate for both groups, and we end up with one result — the difference between the conversion rates.

However, we can never be 100% sure about which population the difference is deriving from — the null or the alternative hypothesis.

There is a way of rejecting the Null hypothesis and conclude that our experiment has an effect — if the probability of observing such difference is relatively small assuming there is no difference.

Hypothesis testing

The last step of our analysis is testing our hypothesis. Since we have a very large sample, we can use the normal approximation for calculating our p-value (i.e. z-test).

p-value is the probability that a sample will have an effect at least as extreme as the effect observed in your sample *if* the null hypothesis is correct. To reject our null hypothesis, we hope that the p-value is small enough.

Calculations return the result:

P-value = 0.00003 < 0.05

We reject the null hypothesis that no difference in the user conversion rate between the control and treatment group.

The scientific standard is to use a p-value less than 0.05, 0.05 here is the significance level (α), meaning that if there is truly no effect, we can correctly infer there is no effect 95 out of 100 time — the result is statistically significance. The equivalent way of accessing statistical significance is to find the confidence interval.

Confidence interval

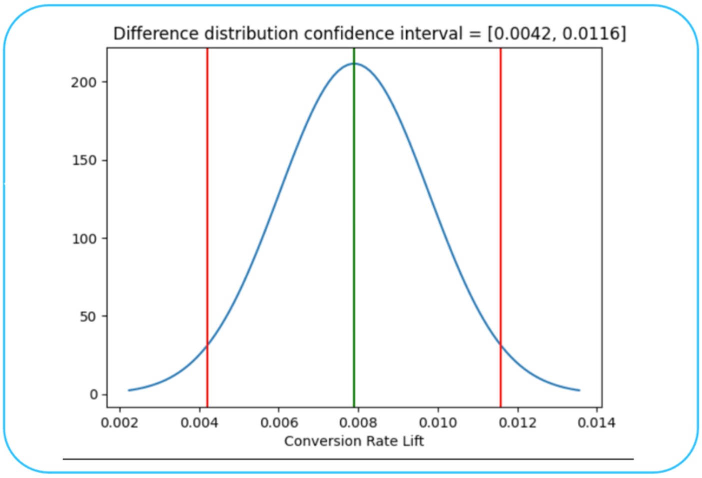
We constructed the 95% confidence interval for the difference in the conversion rates and it lies between 0.42% and 1.16%.

With 95% confidence, the population proportions of users in the treatment group which see new banner is 0.42% to 1.16% higher than the population proportions of users in the control group. In other words, on average, at a 95% confidence level, the conversion rates of the treatment group will be 0.42% - 1.16% higher than the conversion rates of the control group.

The biggest advantage of calculating Confidence Interval for the difference of means is that it tells us how large/small the effect would be between the populations. Reporting only p-value will fail to provide this effect.

Plot a green vertical line at the distributions mean, and a red vertical lines at each of the lower and upper confidence interval bounds.

Difference distribution confidence interval



This means the difference in conversion rate between groups lies in the range of 0.42% - 1.16% for 95% of your entire users. However, only 5% of the time, the difference in conversion rates for these groups will be out of the interval.

In the results, mean difference is 0.79%, and we can be 95% confident that the population difference falls within the range of 0.42% to 1.16%. Here, the Confidence Interval represents a range of values that likely contain the difference between means for the entire population. As the range does not contain 0, results are statistically significant (Zero indicates no difference between the means).

This interval is not too wide and hence, we can be confident that the results will bring meaningful benefits in the future.