"GloBox" A/B Test – Omer Mazor

19/10/2023

Index

Part 1: Data

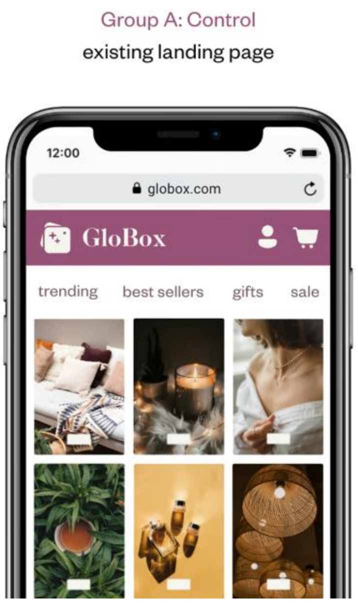
Part 2: Test Statistics

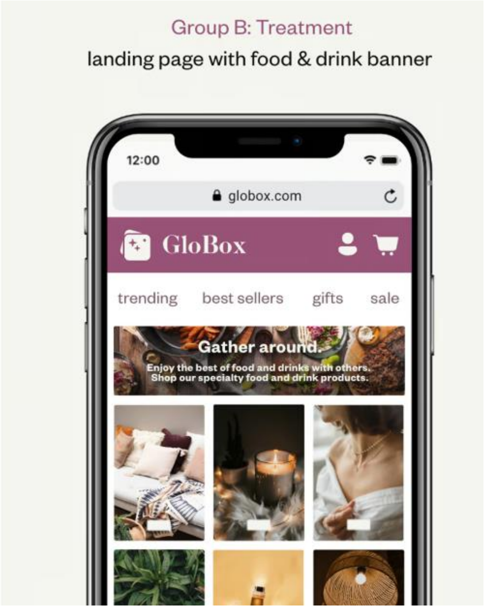
Part 3: Advanced Tasks

Part 4: Summarize and Recommendation

"GloBox" A/B Test Overview

The Growth team decided to run an A/B test highlighting 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 as shown below:




Group B: Treatment
landing page with food & drink banner
12:00
globox.com
GloBox
trending
best sellers
gifts
sale
Gather around.
Enjoy the best of food and drinks with others.
Shop our specialty food and drink products.

In this project, we have to investigate a central question: Is placing the new banner on the site’s main page necessary to increase the company's revenue or not?

Part 1

Let's answer the quiz questions to understand the database

Since we are dealing with data for the purpose of drawing conclusions for the study, we will first want to understand the basic sample data of the experiment.

Can a user show up more than once in the activity table? Yes or no, and why?

Yes, A user can make a purchase more than once on different dates. We will use an SQL query to answer this question formally. We'll group the users by the count of the purchases in descending.

```
select uid, count(uid)

from activity

group by uid

order by count(uid) desc
```

What type of join should we use to join the user's table to the activity table?

We'll use LEFT JOIN to save all the users in the new table because we want to include users regardless of whether they make a purchase.

What SQL function can we use to fill in NULL values?

One of the users of the COALESCE() function is used to replace null entries with a given default value.

What are the start and end dates of the experiment?

25/01/2023 – 06/02/2023. We use an SQL query from the "groups" table to find the first (MIN) and last (MAX) user joining date (join_dt).

```
select min(join_dt), max(join_dt)

from groups
```

How many total users were in the experiment?

48943. A simple SQL query for all records in the "users" table.

```
select *

from users
```

How many users were in the control and treatment groups?

A(Control)- 24343, B(Treatment)- 24600. Counting the number of each group occurrences in the "groups" table and group this counting by the groups.

```
select "group", count("group")

from groups

group by "group"
```

What was the conversion rate of all users?

4.278%. We create a temporary table by joining the "users" and "activity" tables together, adding a column representing whether a user made a purchase or not by 0 and 1, and removing user rows that appear more than once. From the table we created, we select the conversion rate of all users.

```
with isconvert_users as (select distinct id,

case when ac.uid is NULL then 0

else 1

end as is_convert

from users us left join activity ac on us.id =
ac.uid)

select avg(is_convert)

from isconvert_users
```

What is the user conversion rate for the control and treatment groups?

A(Control)- 3.92%, B(Treatment)-4.63%. Now we have to join the "groups" table with the "activity" table with the same structure as before but now group the conversion rate by the groups.

```
with isconvert_users as (select distinct gr.uid,
"group",

case when ac.uid is NULL then 0

else 1

end as is_convert

from "groups" gr left join activity ac on gr.uid =
ac.uid)

select "group", avg(is_convert)

from isconvert_users

group by "group"
```

What is the average amount spent per user for the control and treatment groups, including users who did not convert?

A(Control)-3.374, B(Treatment)-3.390. We create a temporary table for joining the "groups" table with the "activity" table, replace the NULL values with 0, and then calculate the total spent for each user. Then, we use that table to calculate each group’s average spent of all users.

```
with sum_spents as(select gr.uid, "group",

sum(coalesce(spent, 0)) as spent_amount

from "groups" gr left join activity ac on gr.uid =
ac.uid

group by gr.uid)

select "group", avg(spent_amount)

from sum_spents

group by "group"
```

Why does it matter to include users who did not convert when calculating the average amount spent per user?

In order to measure the impact on total revenue (amount spent), we cannot only average the users who converted because there could have been fewer users who converted in the treatment. The average is for all test users, not just those who purchased.

Extracting the analysis dataset:

We'll write a SQL query that returns the user ID, the user’s country, the user’s gender, the user’s device type, the user’s test group, whether they converted or not (spent > $0), and how much they spent in total. Then, we'll download the data as a CSV for the next steps in our research. First, we will select the calculated amount spent for each user in the "activity" table and build it as a temporary table using the "with" clause. Then we will left-join this table to the "group" table (in order to select its "device" and "group" columns) and left-joining "user" table (for its "country" and "gender" columns).

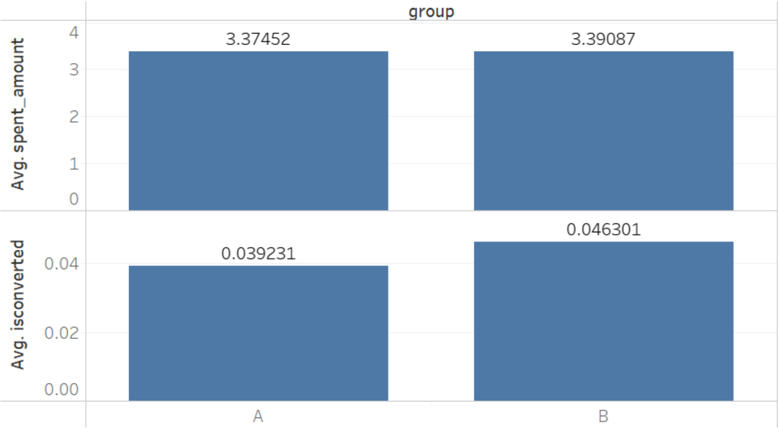
```
with activity_amount_spent as (select uid, sum(spent) as spent
from activity
(group by uid
,"select id, country, gender, gr.device, "group
,case when activity_amount_spent.spent > 0 then 1 else 0 end as isconverted
coalesce(activity_amount_spent.spent, 0) as amount_spent
from users us join groups gr on us.id = gr.uid
left join activity_amount_spent on us.id = activity_amount_spent.uid
```

Visualizations

After we have extracted the data, we will create different charts in Tableau for a clearer illustration of the differences between the groups, the genders, the mobile devices, and the countries in important issues related to purchases on the website.

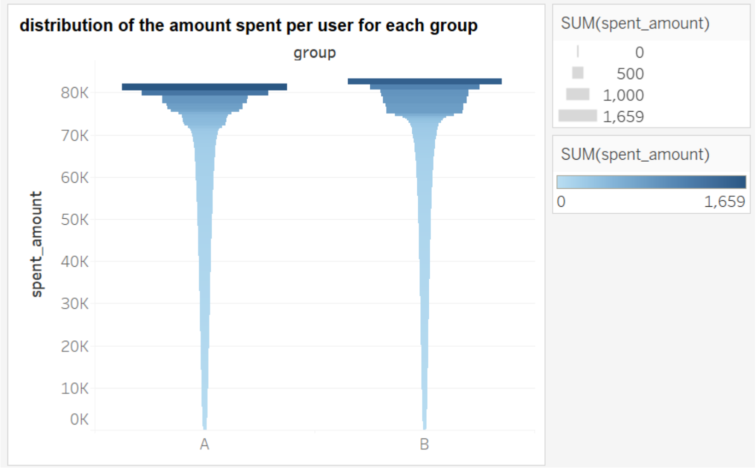
Conversion rate and the average amount spent between the test groups:

Link



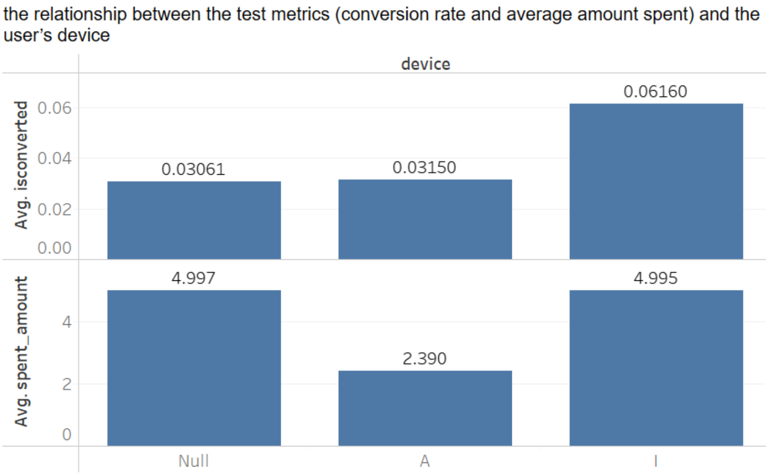
Distribution of the amount spent per user for each group:

Link



The relationship between the test metrics (conversion rate and average amount spent) and the user’s device:

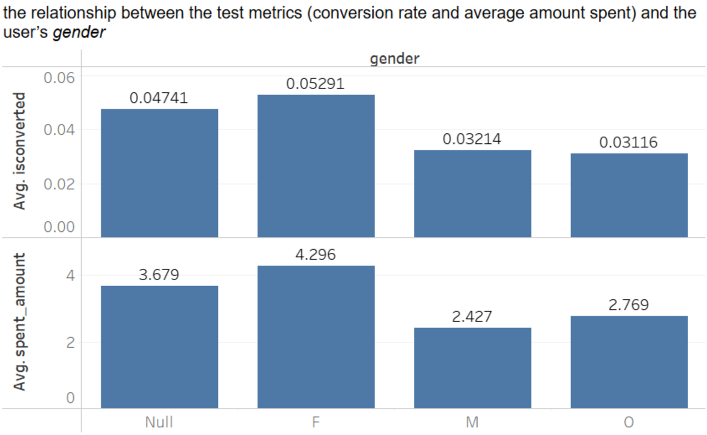
Link



the relationship between the test metrics (conversion rate and average amount spent) and the user’s device

The relationship between the test metrics and the user’s gender:

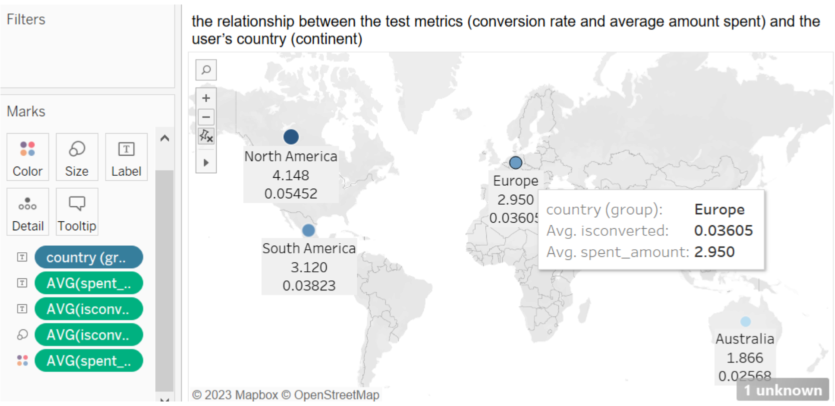
Link



the relationship between the test metrics (conversion rate and average amount spent) and the user’s *gender*

The relationship between the test metrics and the user’s country(continent):

Link



the relationship between the test metrics (conversion rate and average amount spent) and the user’s country (continent)

country (group): Europe
Avg. isconverted: 0.03605
Avg. spent_amount: 2.950

Part 2

Let's calculate A/B test statistics using spreadsheets

For a clearer understanding of the sample results to test the effectiveness of the new banner, we will need to know if there really is a substantial and unequivocal difference between the two groups in the relevant indicators: conversion rate and average spend of each user. We will do this with the help of a hypothesis test and finding the confidence interval.

Conduct a hypothesis test to see whether there is a difference in the conversion rate between the two groups. What is the resulting p-value and conclusion?

We use the normal distribution and a 5% significance level.

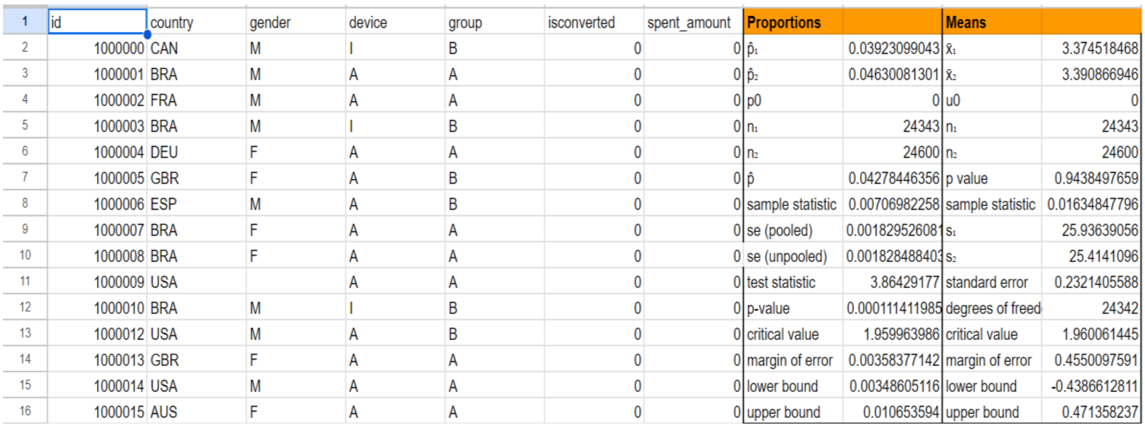
First, we need to determine the two hypotheses about the difference in the conversion rate between the two groups:

The null hypothesis (H0) - the default position, according to which there is no difference between the two measures, meaning that the conversion rate in group A (denoted as p1) is equal to the conversion rate in group B (denoted as p2). Represented as H0: p1 = p2

Alternative hypothesis (H1) - a position that we will try to confirm, which is not the null hypothesis, meaning that the conversion rates between the two groups are different. is represented as H1: p1 !=p2.

Since we are working with a conversion rate that is a percentage and comparing it between two samples, in addition, that is a 'two-tailed' test because the alternative hypothesis claims that the proportion is different (larger or smaller) than in the null hypothesis we'll use two-sample z-test for a difference in proportions.

To calculate the p-value and reach a conclusion, we will have to calculate some data first using Google Sheets (the image below is added to get a little information about the position of each column in the spreadsheet):



1	id	country	gender	device	group	isconverted	spent_amount	Proportions		Means	
2	1000000	CAN	M	I	B	0	0	$\hat{p}_1$	0.03923099043	$\bar{x}_1$	3.374518468
3	1000001	BRA	M	A	A	0	0	$\hat{p}_2$	0.04630081301	$\bar{x}_2$	3.390866946
4	1000002	FRA	M	A	A	0	0	p0	0	u0	0
5	1000003	BRA	M	I	B	0	0	n_1	24343	n_1	24343
6	1000004	DEU	F	A	A	0	0	n_2	24600	n_2	24600
7	1000005	GBR	F	A	B	0	0	$\hat{p}$	0.04278446356	p value	0.9438497659
8	1000006	ESP	M	A	B	0	0	sample statistic	0.00706982258	sample statistic	0.01634847796
9	1000007	BRA	F	A	A	0	0	se (pooled)	0.001829526081	s_1	25.93639056
10	1000008	BRA	F	A	A	0	0	se (unpooled)	0.001828488403	s_2	25.4141096
11	1000009	USA		A	A	0	0	test statistic	3.86429177	standard error	0.2321405588
12	1000010	BRA	M	I	B	0	0	p-value	0.000111411985	degrees of freed	24342
13	1000012	USA	M	A	B	0	0	critical value	1.959963986	critical value	1.960061445
14	1000013	GBR	F	A	A	0	0	margin of error	0.00358377142	margin of error	0.4550097591
15	1000014	USA	M	A	A	0	0	lower bound	0.00348605116	lower bound	-0.4386612811
16	1000015	AUS	F	A	A	0	0	upper bound	0.010653594	upper bound	0.471358237

Link for that spreadsheet

$\hat{p}_1$ - A's conversion rate, we calculate it by the average users that did a conversion (1) or not (0).Formula: AVERAGEIF(E2:E, "A",F2:F). result: 0.0392

$\hat{p}_2$- B's conversion rate. Formula: AVERAGEIF(E2:E, "B",F2:F). Result: 0.0463

n_1- A's group size. Formula: COUNTIF(E2:E,"A"). Result: 24343

n_2- B's group size. Formula: COUNTIF(E2:E,"B"). Result: 24600

$\hat{p}$- the pooled sample proportion: $\frac{\hat{p}_1 \times n_1 + \hat{p}_2 \times n_2}{n_1 + n_2}$.Result: 0.0427

sample statistics - the difference between the two proportions.

standard error - the standard deviation of the estimator for the statistic generated from a sample. We can calculate it with the pooled proportion by this formula: $\sqrt{\hat{p}\ (1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$.Result: 0.001829

test statistic – it's the sample statistic / standard error.Result: 3.864

p-value - The probability of obtaining test results at least as extreme as the result actually observed (our different conversion rates between the groups), under the assumption that the null hypothesis is correct. Formula: 2*NORMSDIST(-I11). Result: 0.0001, statistically significant.

Conclusion: According to the p-value, which is smaller than the established significance level (probability of error, 0.05), we reject the null hypothesis that there is no difference in the user conversion rate between the control and treatment.

What is the 95% confidence interval for the difference in the conversion rate between the treatment and control (treatment-control)?

A 95% confidence interval for the difference in the conversion rate between the treatment and control groups is a statistical range that provides an estimated interval within which the true difference in conversion rates is likely to fall with a confidence level of 95%. In other words, it offers a range of values, and we can be 95% confident that the true difference in conversion rates between the treatment and control groups lies within this interval. This calculation involves sample data and statistical methods to make inferences about the population from which the samples were drawn.

To calculate the 95% confidence interval, we must first know what type of confidence interval we should use.

Like the previous hypothesis test we made, we are working with proportions (conservation rate) and comparing it between two samples, so we should use a Two-sample z-interval for a difference in proportions and calculate its necessary data:

Critical value - threshold value or values that help determine whether to reject the null hypothesis. Formula: NORMSINV(1-((0.05)/2)). Result: 1.959

standard error (unpooled proportions) - $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

margin of error -standard error X critical value. Result: 0.00358

lower bound - sample statistic - margin of error. Result: 0.00348

upper bound - sample statistic + margin of error. Result: 0.01065

Conclusion: we get that the 95% confidence interval is: (0.00348, 0.01065), which means that we can be 95% confident that the true difference in conversion rates between the treatment and control groups lies between 0.00348 and 0.01065.

Conduct a hypothesis test to see whether there is a difference in the average amount spent per user between the two groups. What are the resulting p-value and conclusion?

We use the normal distribution and a 5% significance level.

As we did before, we first determine the null and alternative hypothesis, now about the difference in the average spent per user between the two groups:

The null hypothesis (H0) - the default position, according to which there is no difference between the two measures, meaning that the average amount spent per user in group A (denoted as µ1) is equal to the average amount spent per user in group B (denoted as µ2). Represented as H0: µ1 = µ2

Alternative hypothesis (H1) - a position that we will try to confirm, which is not the null hypothesis, meaning that the average amount spent per user between the two groups IS different. is represented as H1: µ1 != µ2.

Now, unlike the previous test, we are working with an average amount that is actually the mean and comparing it between two samples, so we'll use a two-sample t-test for a difference in means.

To conclude, we first calculate the p-value pertaining to that t-test using Google Sheets:

p-value - The probability of obtaining test results is at least as extreme as the result actually observed (our average amount spent per user between the groups), under the assumption that the null hypothesis is correct. To calculate the p-value in the t-test using Google Sheets, we use the formula T-TEST, which takes 4 parameters: the first parameter is the size of the first group, the second parameter is the size of the second group, the third is the number of the "tails" being tested. Since the alternative hypothesis claims that the means are different (larger or smaller) than the null hypothesis we choose the 'two-tailed' test. The last parameter is the type of the test, which is 2, representing different observations. Formula: =T-TEST(FILTER(G2:G, E2:E = "A"), FILTER(G2:G, E2:E = "B"), 2, 2). Result: 0.9438 statistically insignificant.

Conclusion: According to the p-value, which is greater than the established significance level (0.05), we fail to reject the null hypothesis that there is no difference in the mean amount spent per user between the control and treatment.

What is the 95% confidence interval for the difference in the average amount spent per user between the treatment and the control (treatment-control)?

To know what is the 95% confidence interval, that is, the range of values that we can be 95% sure that the difference in the average amount spent per user between the treatment and control groups is within this interval, we need first know what type of confidence interval we should use?

Like the previous hypothesis test we made before, we are working with means (average amount spent per user) and comparing it between two samples, so we should use a Two-sample t-interval for a difference in means and calculate its necessary data:

$\bar{x}_1$- A's average amount spent per user. Formula: AVERAGE IF(E2:E, "A",G2:G). Result: 3.374

$\bar{x}_2$- B's average amount spent per user. Formula: AVERAGE IF(E2:E, "B",G2:G). Result: 3.39

n_1 and n_2- the size of the groups as we already calculated in the test of proportions.

sample statistic - the difference of the two means.

s_1 - A's standard deviation. Formula: STDEV(FILTER(G2:G,E2:E="A")). Result: 25.936

s_2- B's standard deviation. Formula: STDEV(FILTER(G2:G,E2:E="B")). Result: 25.414

standard error - $\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}$. Result: 0.2321

degrees of freedom – the minimum size between the two groups size subtracted by 1. Result: 24342.

critical value – Formula: T.INV(1-((0.05/2)),K12). Result: 1.96

margin of error - standard error X critical value. Result: 0.455

lower bound - sample statistic - margin of error. Result: -0.438

upper bound - sample statistic + margin of error. Result: 0.471

Conclusion: we get that the 95% confidence interval is (-0.438, 0.471), which means that we can be 95% confident that the true difference in the average amount spent per user between the treatment and control groups lies between -0.438 and 0.471.

Part 3

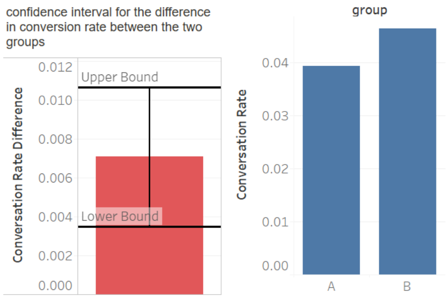
Advanced Tasks

We will add several more tasks to reach a more accurate and clearer conclusion.

Confidence Intervals Visualization

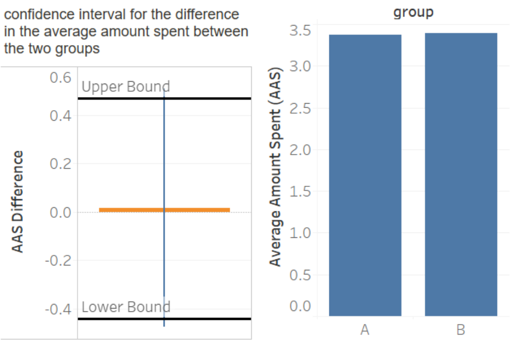
In this task, we will visualize and explain the confidence intervals for the difference in conversion rate and the difference in the average amount spent between the two groups using Tableau. As we explained before, the confidence interval is a range that we can be sure of by a certain percentage (in our experiment, 95%) that the value we are exploring (the difference in the conversion rate and the difference in the average amount spent) is within it.

Link



The image above describes the confidence interval for the difference in the conversion rate between the two groups: the red bar chart on the left symbolizes the difference we found in our sample (the difference in conversion rate between the two groups, right side image) in favor of group B and its value is 0.007070.

The upper black line denotes the upper bound of the confidence interval, and the lower black line denotes its lower bound. The line between them is the range's length or size (the bounds difference). The upper limit is 0.01065, and the lower limit is 0.00348. It means that if we take an infinite number of samples, with a 95% chance, the difference in conversion rates between the two groups will be at least 0.00348 and at most 0.01065 in favor of group B.

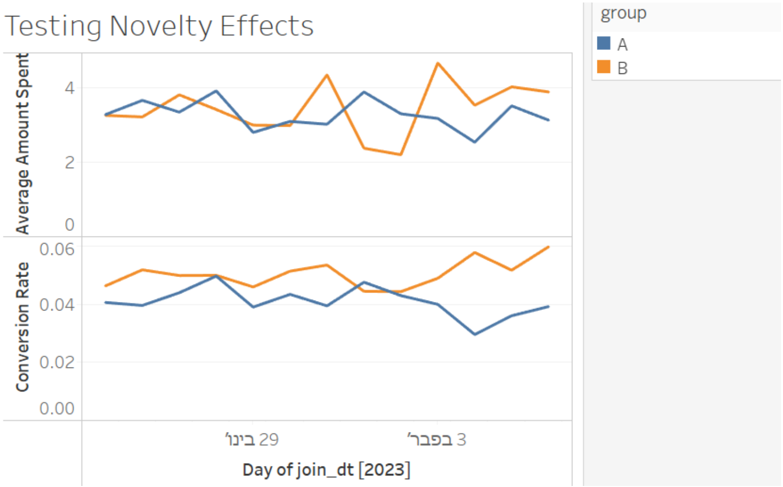


The image above describes the confidence interval for the difference in the average amount spent between the two groups: the thin orange bar chart on the left symbolizes the difference we found in our sample (the difference in average amount spent between the two groups, right side image) in favor of group B and its rate is 0.01065.

The upper black line denotes the upper bound of the confidence interval, and the lower black line denotes its lower bound. The line between them is the length or the size of the range (the bounds difference). The upper limit is 0.4713, and the lower limit is -0.4386. It means that if we take an infinite number of samples, with a 95% chance, the difference in the average amount spent between the two groups will be at most 0.4386 in favor of group A and at most 0.4713 in favor of group B.

Novelty Effects

LInk



By looking at those charts, we see that, in general, there are no novelty effects. That is, there is no sharp increase in the key metrics in group B at the beginning of the experiment that faded over time, but rather the metrics change unevenly. So we, can't say that the effectiveness of the banner is short-lived.

Power Analysis

Our sample size is 48943, to know what sample size will ensure a high probability that we correctly reject the Null Hypothesis that there is no difference between the two groups in conversion rate and in the average amount spent per user, we need to use the "Power Analysis" method. Power is affected by several things. However, there are two main factors:

1. How much overlap there is between the two distributions we want to identify with our study (conversion rate, average amount spent).

2. The sample size is the number of measurements we collected from each group.

The first thing we need to decide is how much Power (the probability that we'll correctly reject the Null Hypothesis) we want. Although we can pick any value between 0 and 1 for Power, a common value is 0.8. the second thing we need to do is determine the level of significance (probability of error), whose common value is 0.05, so we'll use that.

We'll find the sample sizes for both the difference in conversion rate and the average amount spent:

Sample size for conversion rate:

Using the Statsig Sample Size Calculator for Conversions with a minimum detectable effect of 10%, we get that we need 155,400 measurements per group to reject the Null Hypothesis with 80% chance correctly.

Sample size for average amount:

We need to estimate the overlap between the two distributions. Overlap is affected by the distance between the population means and the standard deviations. A common way to combine the distance between the means and the standard deviations into a single metric is to calculate an "Effect Size," which is also called d: Effect Size(d) = $\frac{The\ estimated\ difference\ berween\ the\ means}{Pooled\ estimated\ standard\ deviations}$. we'll calculate the pooled estimated standard deviations in one of the simplest ways = $\sqrt{\frac{s_1{}^2 + s_2{}^2}{2}}$ when s_1 and s_2 are the standard deviations of the A and B groups respectively that we calculated before with Google Sheets. The result for this is 25.6765. After we calculated those values, we'll use the Statulator Sample Size Calculator for Means to find our desired sample size. The result is 38,721,958. This means that if we get 38,721,958 measurements per group, we'll have an 80% chance that we'll correctly reject the Null Hypothesis.

We found before (Part 2) that for the statistical significance we determine (0.05), we have to reject the null hypothesis that there is no difference in the conversion rate between the control and treatment groups. This means that we have found that this difference comes from not just random chance. But is this change useful from a practical point of view?

Let's look at several important points:

Business impact - according to the general overview of the project, the goal of the central company is to increase revenue, and for that purpose, all this change was made. For this purpose, you can look at the average amount spent for each user. We found before that for the statistical significance we defined, we could not reject the null hypothesis that there is no difference in the average amount spent per user between the control and the treatment. In addition, the difference in our sample was minimal and insignificant, so in practice, it is difficult to determine that the company's goal was achieved following the new banner on the main page.

Consideration of costs and benefits - even if we verify that there has been a certain increase in the amount of revenue since the introduction of the new banner to the main page, we will ask the questions: Is there a maintenance cost for the banner (following its replacement according to the key products in the food and drink category)? If so, what is the relationship between it and the profit (which is minimum, as mentioned)? Would it have been possible to utilize that high-value page space for other, more economically efficient ideas?

Part 4

Summarize our findings from the A/B test results

Let's collect everything together for a summary.

Our A/B test recommendation

Glossary

Null Hypothesis: the hypothesis that there is no difference between the measures.

Confidence interval: estimated range

Significance level: probability of error

We saw that both in the conversion rate and in the average amount spent by each user, the ratio is in favor of group B (see the difference between the groups visualization). This is the test group whose users were exposed to the new banner. For the conversion rate, we found that we are allowed to reject the null hypothesis-the hypothesis that there is no difference (see conclusion of hypothesis test for conversion rate), but we would need 155,400 participants in each group to be 80% confident in this rejection (see sample size for conversion rate). In addition, we found out using the confidence interval that we can be 95% sure that after an infinite number of experiments, the difference between the groups will still be in favor of group B even if it could be minimal (see confidence intervals visualization). Conversely, for the average amount spent by each user in the two different groups, we failed to reject the null hypothesis that there is no difference between the two groups (see conclusion of hypothesis test for average amount spent), and we saw that we would need 38,721,958 participants for each group to be 80% confident in rejecting it (see sample size for average amount). Using the confidence interval, we found that we can be 95% sure that after an infinite number of experiments, the difference between the groups is not clear-cut and could also be in favor of group A, although there is a small advantage for group B following the current sample we performed (see confidence intervals visualization). In conclusion, in terms of the conversion rate, there is a distinct advantage for group B, and for the average amount spent by each user, there is an extremely minimal advantage to the point that it is not possible to be 100% sure for group B. We see a statistically significant higher conversion rate and, in addition, an unequivocal confidence interval. That’s a good start to getting more paying customers. Maybe we can draw them back to make more purchases later. The average amount each user spends in group B is also a little higher but statistically insignificant. The company’s main goal is to increase its revenue, and in our sample it seems that it did increase, albeit slightly, compared to the control group, considering that the confidence interval doesn’t give us significant information about an infinite number of experiences.

Recommendation: Continue iterating. Considering the sample results and the data we collected, I would say that in light of the company's main goal, which is to increase revenue, it is worthwhile to examine the effect of the banner over a longer period and for a higher number of users for our test will be more powerful. At the same time, we should pay attention to the expenses involved in the existence of the banner: a banner is not typically an expensive feature to launch in terms of engineering time or operational overhead, but using the important space on the main page for it at the expense of other content that can be more profitable may be a wrong choice.