Pacmann

Marketing of Public Service Announcement A/B Testing

Experimental Design and A/B Testing

Outline

- Introduction/Background
- Topics, Datasets, and Tool
- Define Problem, Goals, Metrics
- Dashboard User Flow
- Data Cleaning
- Dashboard and Business Insights
- Conclusion, Solution
- References

Introduction

- Hello and welcome to my Business Intelligence portfolio! My name is Mory Handy. I am just starting my journey in the field of data and eager to learn and grow as a professional. I have completed several online courses from Pacmann and have hands-on experience with data cleaning, visualization, and simple statistical analysis. My goal is to continue learning and expanding my skill set, and this portfolio showcases the projects I have worked on so far. I hope that this portfolio gives you a glimpse into my enthusiasm for data and my ability to learn quickly and apply new techniques to real-world case. Thank you for visiting my portfolio, hope all of you can get insight from this case

- Medium : https://medium.com/@mrmorry77
- Github : https://github.com/MoryHandy13

Background

A public service announcement (PSA) is a message created and distributed in the media to inform and educate the public about a particular issue or topic. PSAs are typically sponsored by nonprofit organizations, government agencies, or other public interest groups and are intended to raise awareness, change attitudes, and encourage positive behaviors.

Marketing companies want to run successful campaigns, but the market is complex and several options can work. So normally they tun A/B tests, that is a randomized experimentation process wherein two or more versions of a variable (web page, page element, banner, etc.) are shown to different segments of people at the same time to determine which version leaves the maximum impact and drive business metrics.

Dataset

The goals of the dataset is to analyze the groups, find if the ads were successful, how much the company can make from the ads, and if the difference between the groups is statistically significant.

Data dictionary:

- Index: Row index
- user id: User ID (unique)
- test group: If "AD" the person saw the advertisement, if “PSA” they only saw the public service announcement
- converted: If a person bought the product then True, else is False
- total ads: Amount of ads seen by person
- most ads day: Day that the person saw the biggest amount of ads
- most ads hour: Hour of day that the person saw the biggest amount of ads

Case Problem

Goals

The goals for AB testing on marketing of Public Service Announcement and ad could be to determine which advertising approach is more effective in increasing brand awareness and driving conversions. Specifically, the goal could be to identify which version of the advertising campaign generates higher click-through rates, conversion rates, and ultimately, higher ROI.

To achieve all of this, the experiment could be designed to compare two versions of the same advertising campaign (A and B), with one version featuring a Public Service Announcement approach and the other version featuring a traditional ad approach. The experiment could then measure the effectiveness of each approach by tracking metrics such as click-through rates, conversion rates, and revenue generated. The ultimate goal would be to identify which approach generates the most positive results, and to use that information to inform future advertising strategies.

Metrics

Driver Metrics should be in line with the goal metric, sensitive, actionable and can be meaningful by short-term experiment. Those characteristics lead the metrics fit with the experiment.

Guardrail metrics give an alert about outcome of the experiment that potentially misleading. These metrics monitor the trade-off that undesirably happen. Also can be use for sanity check about outcome experiment.

There are 2 kind of guardrail metrics:

a. Organizational Guardrail Metrics
    - To see if there any others trade-off happen when running the initiative
    - If these metrics lead to negative impact, the business can lead to losses
b. Trust-Related Guardrail Metrics
    - Monitoring level of confidence (trustworthiness) of the experiment
    - Checking the infraction of assumptions

Metrics

Objective	How to Achieve	Intended Outcomes	Driver Metrics	Guardrail Metric
Increase revenue of marketing company	Increase user subscribe Use ads that leads to conversion	Increase brand awareness and driving conversions	Conversion rates	SRM (Sample Ratio Mismatch)

To determine the driver metrics, the metrics should be use to monitor the behavior of the user from data collected (measurable). This metrics also can be used to measure the effect of initiave from variant control & treatment (Attributable). Driver metric is leading indicator from goal metrics. So this metrics should have enough variability that can differentiate treatment and control (Sensitivity). Last, the metrics should can be measure by short-term (Timely). By all of those characteristics, writers choose Conversion Rates become Driver Metrics of this AB Testing cases.

Variants

Control : Public Service Announcements (PSA)
Treatment : Creative Ads, such as
a.Ads with a good headline
b.Make the text of information in ads more concise and informative
Good CTA button that can invite customers to engange.
E.g : “Get Now!”, or “Buy Now!”

The experiment using total ads below and above 15 based on median of total ads each customer.

Variants	Ads	Total Ads	Keterangan
A	PSA	<15	Control
B	PSA	>15	Treatment 1
C	Creative Ads	<15	Treatment 2
D	Creative Ads	>15	Treatment 3

Hypothesis

Goal : See the ad impact towards the conversion rates

We want to compare whether group i th is more than group j th, so we use one sided (right tail) hypothesis testing.

We want to prove whether the conversion rate of group j is greater than the conversion rate of group i

Hypothesis :

- group A vs group B

$$H_0 : p_B \leq p_A$$
$$H_1 : p_B > p_A$$

- group A vs group C

$$H_0 : p_C \leq p_A$$
$$H_1 : p_C > p_A$$

- group A vs group D

$$H_0 : p_D \leq p_A$$
$$H_1 : p_D > p_A$$

- group B vs group C

$$H_0 : p_C \leq p_B$$
$$H_1 : p_C > p_B$$

- group B vs group D

$$H_0 : p_D \leq p_B$$
$$H_1 : p_D > p_B$$

- group C vs group D

$$H_0 : p_D \leq p_C$$
$$H_1 : p_D > p_C$$

Experimental

Experimental

- **Randomization Unit**
  Randomization unit is "who" or "what" kind of thinks that allocated randomly to each group. To get more context of the experiment, we're limiting the population of people in Jakarta.
- **Target of randomization unit**
  Target of randomization unit is all user that exposed by Public Service Announcement and Creative Ads in Jakarta. Also, we would considering about total ads that expose to the customer as variant of the randomization unit

- **Sample size**
  Size of sample will affect the power of evidence to show validity the experiment.
  a. Significant level (α)
     α = P (Accept H1 | Ho right)
     That means opportunities to accept H1, whereas Ho right. Because those things is wrong, so we should reduce the α value. Conservatively, industry rules using 5% or 1% for α values. We're determined to use α = 5% as significant level.

Experimental

- **Sample size**

  b. Power level ($1 - \beta$)

  $1 - \beta = P$ (Reject Ho | Ho wrong)

  That means opportunities to reject Ho, whereas Ho wrong. Because those things is right, so we should increase the $1 - \beta$ value. Conservatively, industry rules using 80% for power lever. We're determined to use $1 - \beta = 80\%$ as power level.

  c. Standard deviation of population (σ)

  For this experiment, we make an assumption of standard deviation population is 0.1

  d. Difference between control and treatment (δ)

  For business propose, we make an assumption that these treatment will be profitable if the conversion rate increase 1%. So, the management will be implemented the Creative ads rather than PSA because the impact of increasing conversion rate.

Experimental

- **Calculating Sample size**
  Remember, we can use the given formula to calculate the minimum number of sample size needed.

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

  Thus, if we have the z value, we can determine the number of sample. Using equation above, we will get **1.570 user** sample needed. Also we can determine total sample for four group is **6.280 user.**

  In that case, we don't know the standard deviation of the conversion rate. However, we can calculate the standard deviation with the information of current baseline conversion rate.
  The conversion event is a Bernoulli trial, with p = 0.02. We can calculate the standard deviation by using the following formula with approach of Bernoulli distribution:

$$\sigma = \sqrt{\hat{p}(1-\hat{p})}$$

  With the equation above, we get number of sample needed by **3.080** user. Because standard deviation is higher, so the sample size is higher. And we need **12.320** sample for four groups.

Experimental

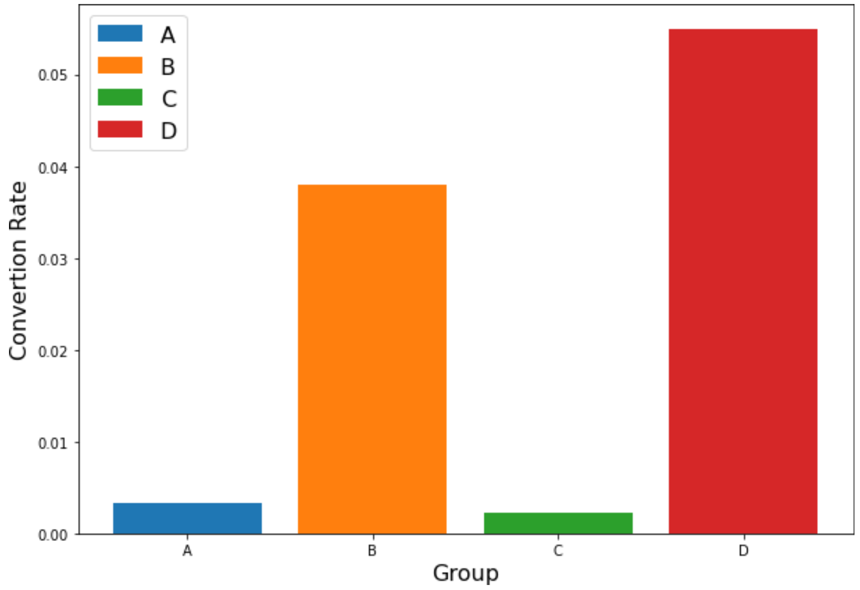
- **How long run experiment**
  For make an assumption that every week we can gain 3.000 samples, so we can calculate the experiment time is 5 weeks

Experiment and Data Gain

Experiment and Data Gain

We can't gain data directly, we're assume the following dataset is our experiment. We can take a sample according to the experimental design we made by Designing Experiment.

From **dataset marketing_AB.csv**, we get 588.101 row and with Simple Random Sampling, we take 3.080 user as experiment data per variant or totally 12.320 sample. Then we create group of variant as the feature . We calculate the number of percentages Conversion Rates by the random sampling as it is



Group	#User	#Convert	Convertion Rate
A	3080	14	0.004
B	3080	104	0.037
C	3080	8	0.004
D	3080	173	0.053

Analyze The Data

Ensure the Quality of Data

a. **Check the data quality (missing value, duplicate data, distribution of data)**

Some mechanism to ensure trustworthiness are:

- Validate data quality
- Avoid threat to internal validity
- Avoid threat to external validity
- Mitigate the effect of simpson's paradox

Data Quality

We can use the following checklist to measure data quality :

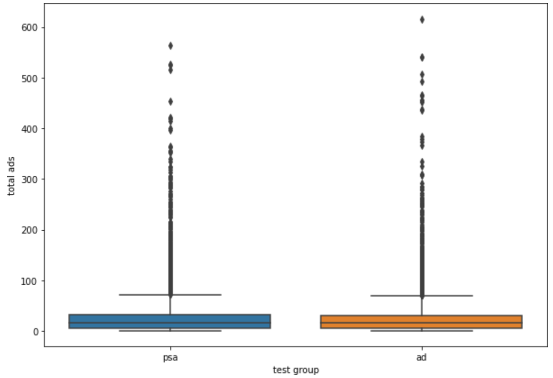
- Missing rates : How much missing value in dataset
- Uniqueness : No duplicate data
- Invalid values : Do the values follow the proper format? Are the values valid for the variable/column?
- Data delays : How many data is there at the periode of the experiment? How long does it take between when the events were logged and when the data is available for analysis?

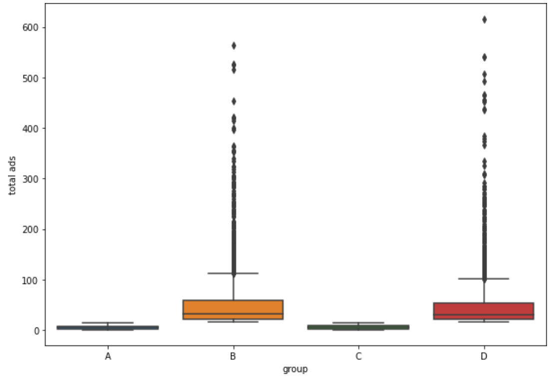
We're also do the checking for any NaN data (Missing value), Duplicate Data, and Invalid Data by it own combination

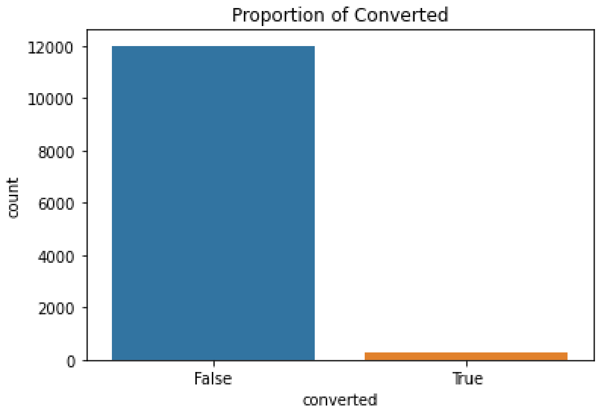
Ensure the Quality of Data

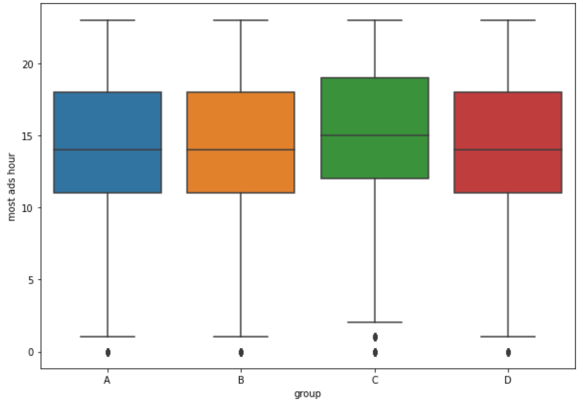
b. **Data exploration** (how many users in each group, and other insight from dataset that has been choose)

- From data exploration, we can see that there are 3.080 user with 25% percentage for each group
- Using box plot, we can analysis there are no difference between total ads that consume by customer with psa and creative ads
- There are disparities of proportion between converted user in marketing funnel. And there are not significant difference between hour of ad between all variants





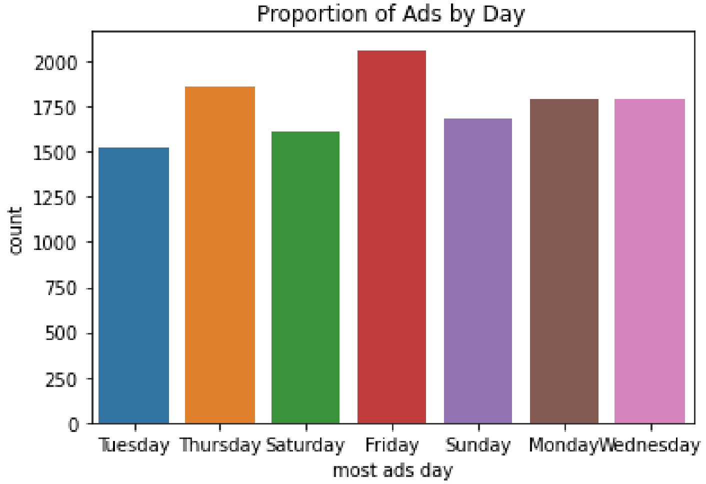


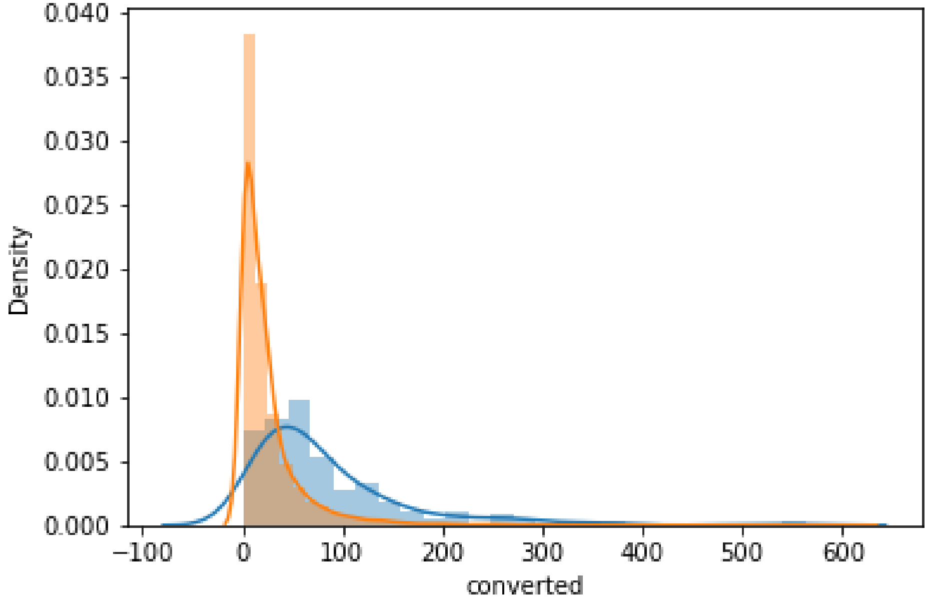


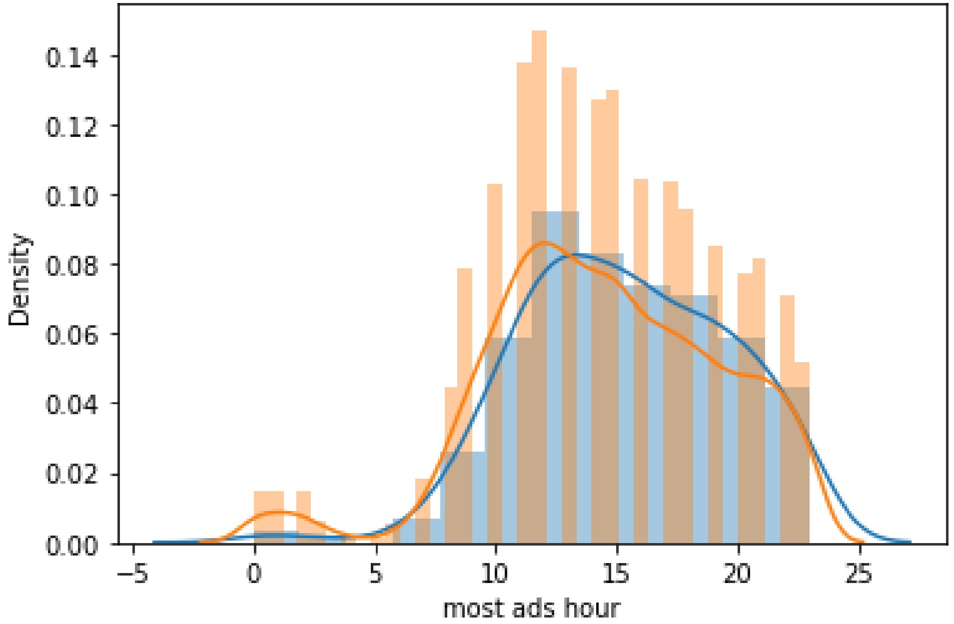
Ensure the Quality of Data

b. **Data exploration** (how many users in each group, and other insight from dataset that has been choose)

- Ads and psa has much implemented at Friday, and has less implementation at Tuesday ada Saturday.
- Using distribution plotting, we can distinguish between converted and unconverted customers based on their ads that has been consumed. There are significant difference of distribution. From most ads hour, as we see there aren't any significant differences both of distribution







Pacmann

Ensure the Quality of Data

c. Perform SRM test with chi-square test

Sample Ratio Mismatch (SRM) is the situation when the observed sample ratio in the experiment is different from the expected.

Chi-square test can be used to detect whether an experiment has SRM or not.

The steps for doing a chi-square test in order to detect SRM are:

1. Define the null and alternative hypothesis (H0 and H1)
2. Calculate chi-square statistics
3. Define decision rules
4. Make decisions and draw a conclusion

The steps for doing a chi-square test in order to detect SRM are:

1. Define the null and alternative hypothesis (H_0 and H_1)

H_0 : No SRM detected

H_1 : SRM detected

2. Calculate chi-square statistics

$$\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

Where :

Observed: the control and variation traffic volumes (sample size), respectively

Expected: the expected values for control and treatment — i.e. the total observed divided by 2

Observed is the same as # user in each group.

For calculate expected in each group, we can use total observed divided by 4

Ensure the Quality of Data

c. Perform SRM test with chi-square test

Define decision rule

In making statistical test decisions, we can use:

- Comparison of chi-square statistics with critical value
  - $\chi^2 > \chi^2_{\alpha,df} \rightarrow$ reject H_0
- Comparison of p-value with alpha
  - pvalue < $\alpha \rightarrow$ reject H_0

Normally, one would look for a p-value of 0.05 or less to proof of SRM. The problem with 0.05 is that it's not strict enough for our purposes. Using this might give us a false signal. What we need is to be stricter for our test. So we use significance level 1%.

degree of freedom (df) is calculated as:

$$df = (rows - 1) \times (columns - 1)$$

Ensure the Quality of Data

c. Perform SRM test with chi-square test

Comparison of chi-square statistics with critical value. We must calculate the critical first. Critical value is the chi-square value at alpha. And we get critical value 6.635. Make decisions from chi-square statistics and critical value and we calculate **Fail to Reject Ho / No SRM**
Based on data quality, we have done data cleaning so that the data we use is of sufficient quality. But we need to check again, whether the sample size after data cleaning is sufficient (according to the experimental design) or not so that there is enough power to draw credible conclusions.

Based on the detection of SRM, although the sample size of the cleaned data in the control and treatment groups is different. However, SRM was not detected.

Hypothesis Testing and Analyze Data

After running the experiment, we can calculate the lift over baseline by this equation:

$$\text{Lift} = CVR_{\text{treatment}} - CVR_{\text{control}}$$

Lift-over-baseline for treatment B is 3.48 %
Lift-over-baseline for treatment C is -0.09 %
Lift-over-baseline for treatment D is 5.17 %
By this data, we can inference that Treatment D has biggest **lift-over-baseline**

Because there are more than two variants, we do the multiple hypothesis with Benjamini-Hochberg Correction. To find out which one is the best, we can do a hypothesis testing. A suitable hypothesis test for this case is the z-test for proport ion. Because we have more than two groups to compare, therefore we perform multiple hypothesis testing for each gro up pair.

An issue with multiple hypothesis testing is increasing of Type I error, so we can do correction with Benjamini-Hochberg Correction.

Hypothesis Testing and Analyze Data

The following is the stage for conducting the analysis :

a. **Define null hypothesis and alternative hypothesis**

   We want to compare whether group i th is more than group j th, so we use one sided (right tail) hypothesis testing.

   We want to prove whether the conversion rate of group j is greater than the conversion rate of group i

- group A vs group B

$$H_0 : p_B \leq p_A$$
$$H_1 : p_B > p_A$$

- group A vs group C

$$H_0 : p_C \leq p_A$$
$$H_1 : p_C > p_A$$

- group A vs group D

$$H_0 : p_D \leq p_A$$
$$H_1 : p_D > p_A$$

- group B vs group C

$$H_0 : p_C \leq p_B$$
$$H_1 : p_C > p_B$$

- group B vs group D

$$H_0 : p_D \leq p_B$$
$$H_1 : p_D > p_B$$

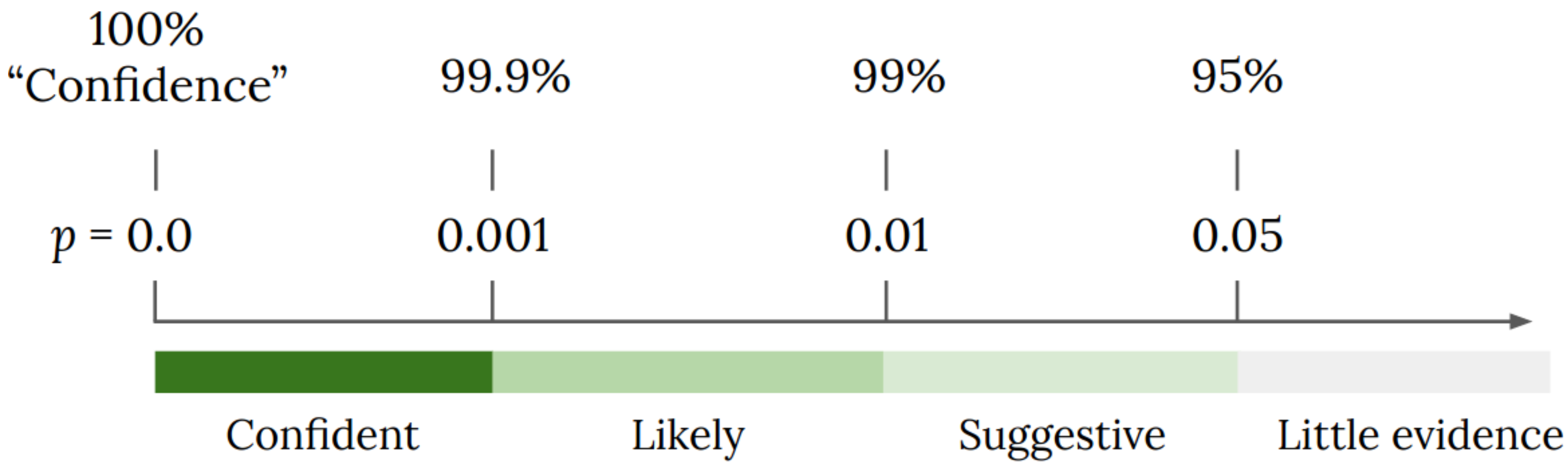
- group C vs group D

$$H_0 : p_D \leq p_C$$
$$H_1 : p_D > p_C$$

And then set siginificance level (alpha) = 0.05

Hypothesis Testing and Analyze Data

b. Calculate the p-value in each test



	pair_group	p-value
0	B > A	4.231391e-22
1	C > A	7.668808e-01
2	D > A	8.512005e-34
3	C > B	1.000000e+00
4	D > B	8.197395e-04
5	D > C	1.489758e-35

Hypothesis Testing and Analyze Data

c. Arrange the p-values in order from smallest to largest (ascending order)

	pair_group	p-value
5	D > C	1.489758e-35
2	D > A	8.512005e-34
0	B > A	4.231391e-22
4	D > B	8.197395e-04
1	C > A	7.668808e-01
3	C > B	1.000000e+00

d. Assign ranks to the ordered p-values

	pair_group	p-value	rank
5	D > C	1.489758e-35	1
2	D > A	8.512005e-34	2
0	B > A	4.231391e-22	3
4	D > B	8.197395e-04	4
1	C > A	7.668808e-01	5
3	C > B	1.000000e+00	6

Hypothesis Testing and Analyze Data

e. Calculate each individual p-value's Benjamini-Hochberg critical value

Using the formula:

$$BH - critical\ value = \left(\frac{i}{m}\right) Q$$

where:

- i = the p-value's rank
- m = total number of tests
- Q = the false discovery rate (chosen by the experimenter)

Suppose that the experimenter want to control false discovery rate in 5%. So the Q = 0.05

	pair_group	p-value	rank	BH-crit
5	D > C	1.489758e-35	1	0.008333
2	D > A	8.512005e-34	2	0.016667
0	B > A	4.231391e-22	3	0.025000
4	D > B	8.197395e-04	4	0.033333
1	C > A	7.668808e-01	5	0.041667
3	C > B	1.000000e+00	6	0.050000

Hypothesis Testing and Analyze Data

e. Calculate each individual p-value's Benjamini-Hochberg critical value

Using the formula:

$$BH - critical\ value = \left(\frac{i}{m}\right) Q$$

where:

- i = the p-value's rank
- m = total number of tests
- Q = the false discovery rate (chosen by the experimenter)

Suppose that the experimenter want to control false discovery rate in 5%. So the Q = 0.05

	pair_group	p-value	rank	BH-crit
5	D > C	1.489758e-35	1	0.008333
2	D > A	8.512005e-34	2	0.016667
0	B > A	4.231391e-22	3	0.025000
4	D > B	8.197395e-04	4	0.033333
1	C > A	7.668808e-01	5	0.041667
3	C > B	1.000000e+00	6	0.050000

Pacmann

Hypothesis Testing and Analyze Data

f. Compare original p-values to the Benjamini-Hochberg critical value

If the original p-values smaller than Benjamini-Hochberg critical, then the test are significant (reject Ho)

	pair_group	p-value	rank	BH-crit	Significant?
5	D > C	1.489758e-35	1	0.008333	Yes
2	D > A	8.512005e-34	2	0.016667	Yes
0	B > A	4.231391e-22	3	0.025000	Yes
4	D > B	8.197395e-04	4	0.033333	Yes
1	C > A	7.668808e-01	5	0.041667	No
3	C > B	1.000000e+00	6	0.050000	No

g. Conclusion

- Based on the results of multiple testing, test for group D vs group C, group D vs group A, group B vs group A, and group B vs group D resulted a significant outcome.
- Because our hypothesis is to compare whether group j is more than group i, so it can be concluded that there is sufficient evidence that the conversion rate of group D (ad + total ads > 15) is higher than groups A, B and C.
- Group D is the group that has the highest conversion rate among all groups.
- Group D becomes the winning version of the 4 combinations of the marketing company.
- It means that using creative ads with total ads > 15 statistically has an impact on increasing conversion rates.

Interval of Difference between Treatment and Control

After that, we will calculate the confidence interval to estimate within what range the difference or proportion discrepancy in the population lies.

	lower	upper
confidence_interval_AB	0.027974	0.042235
confidence_interval_AC	-0.003935	0.001856
confidence_interval_AD	0.043644	0.060361
confidence_interval_BC	-0.043154	-0.029066
confidence_interval_BD	0.006383	0.027492
confidence_interval_CD	0.044717	0.061288

Based on these results, we are 95% confident that the difference in proportion of users who converted between the treatment group (B) and the control group (A) can be seen in the table below. Or it can be said that the increase in conversion rate using the Creative Ad method (treatment) has increased according to the table below.

Recommendation for the marketing company: based on the statistical test results, it is statistically significant. However, to make a decision whether to add the voucher code feature or not, it needs to be ensured whether it is practically significant such as the cost of using ads, marketing costs, etc. should not incur losses.

With a minimum difference in conversion rate of 1%, the Confidence Interval values for **A vs D (4.47% - 6.13%)** so that the use of Creative Ads is recommended.

Probability

To get sense of chance of a variation to have the best performance in the long term, we simulate its probability distribution given the current data Bayesian approach it is. Simply, we use the Bayesian theorem to find our update believe (posterior) about something that we know (prior) given the data (likelihood).

$$P(\mu|z) \text{ a } P(\mu)P(z|\mu)$$

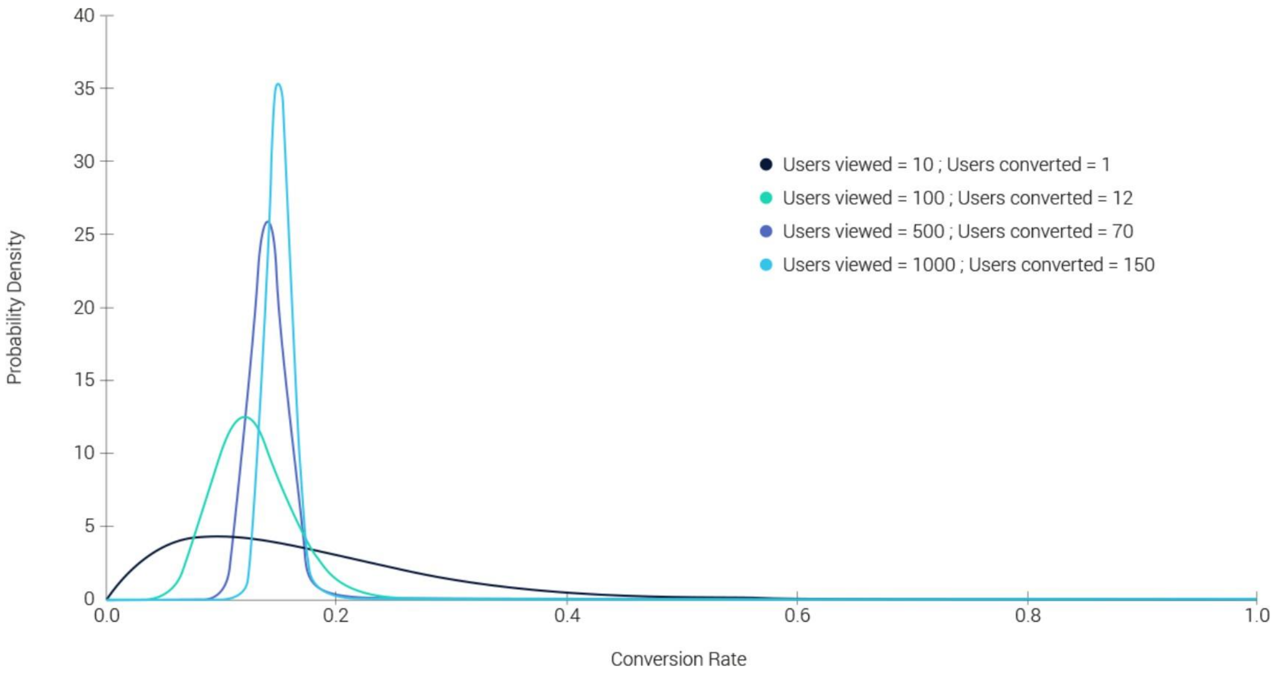
$P(\mu)$ is the prior probability to find our current conversion rate (μ). Because it is convert or not convert, the probability must be following the binomial distribution. Thus,

$$P(\mu) \sim Binomial\ (\mu, n_{trial}, n_{success})$$

$P(\mu|z)$ is the likelihood. Why we need this? Because, even the CVR is similar, however it is difference between:

- 1 conversion from 10 users
- 12 conversion from 100 users
- 70 conversion from 500 users

Probability



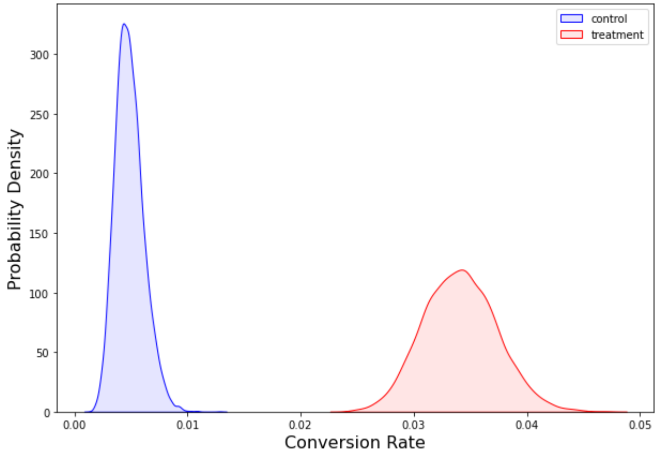
We can model the above distribution using Beta distribution. Why? Because beta distribution return value between 0-1.

$$P(z|\mu) \sim Beta(\alpha|\beta)$$

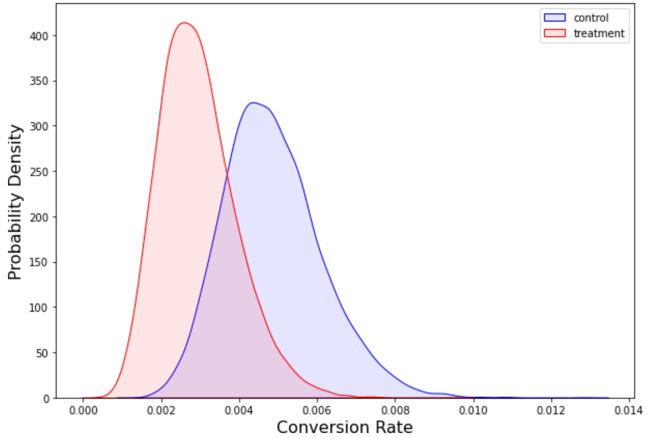
Multiply both the prior and likelihood to obtain the posterior. In short, we got the posterior distribution as

$$P(\mu|z) \sim Beta(\alpha = n_{success} + 1, \beta = n_{fail} + 1)$$

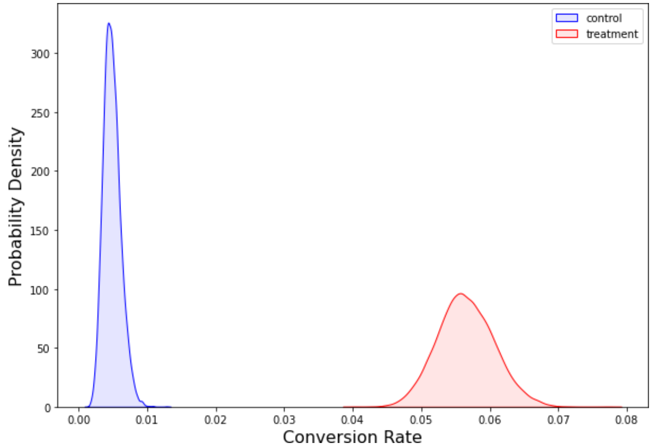
Probability



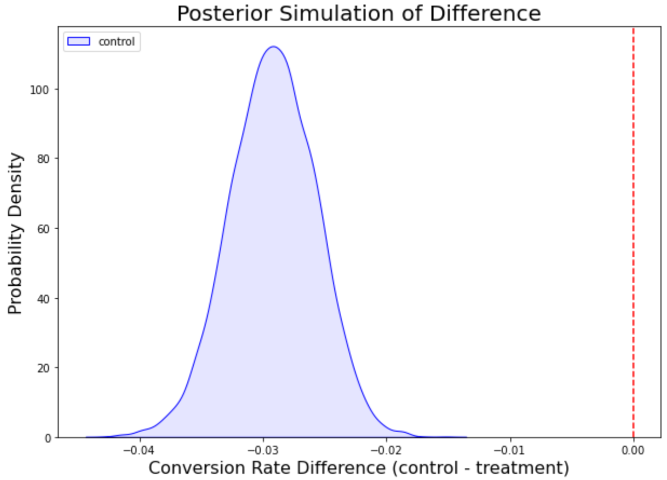
Control Posterior A vs Treatment Posterior B



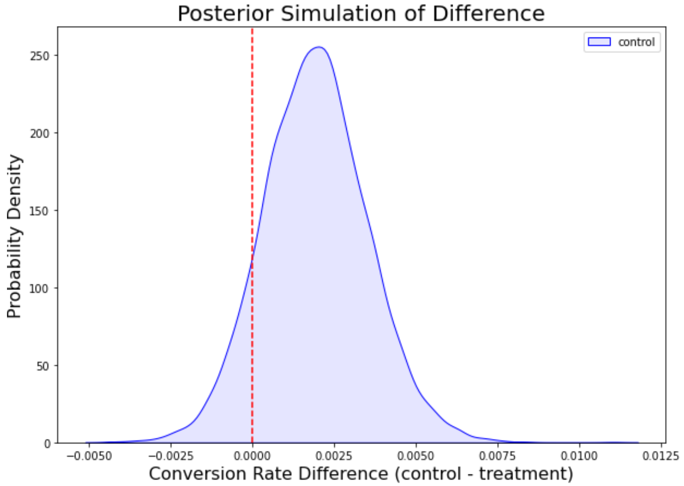
Control Posterior A vs Treatment Posterior C



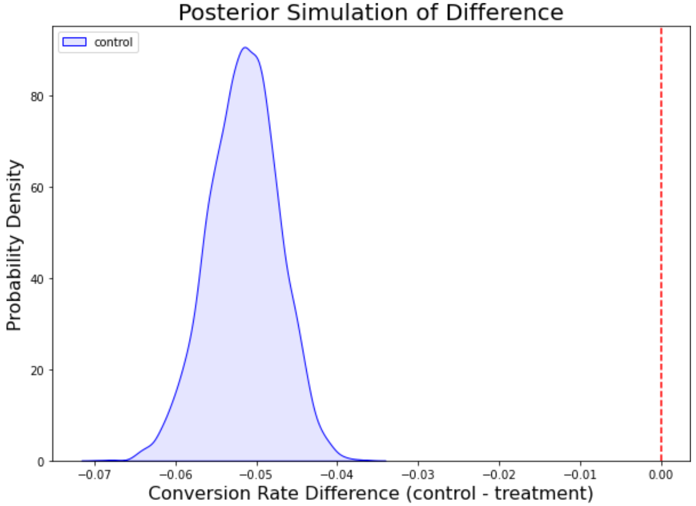
Control Posterior A vs Treatment Posterior D



Control Posterior A vs Treatment Posterior B

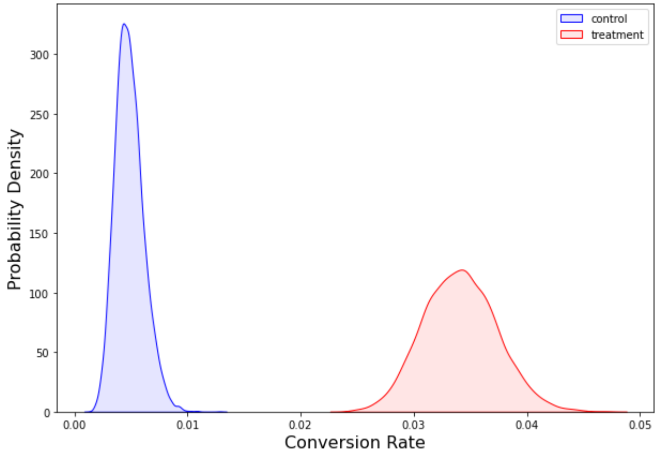


Control Posterior A vs Treatment Posterior C

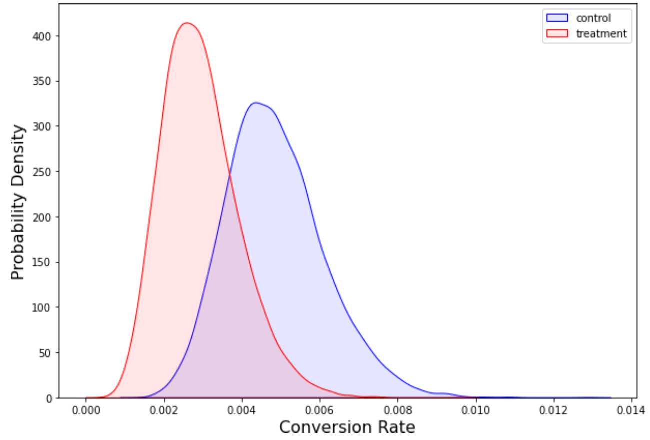


Control Posterior A vs Treatment Posterior D

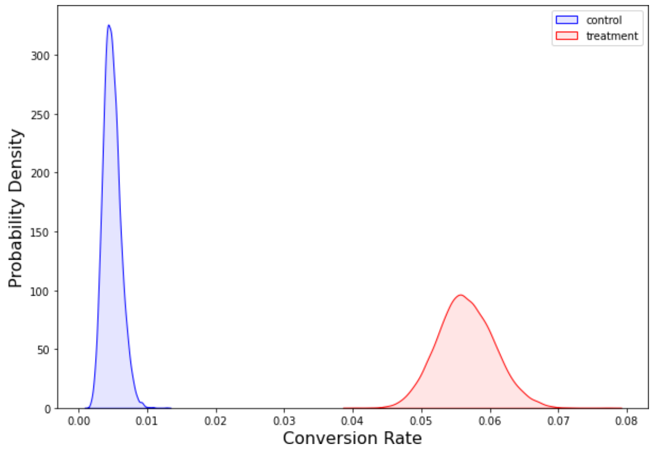
Probability



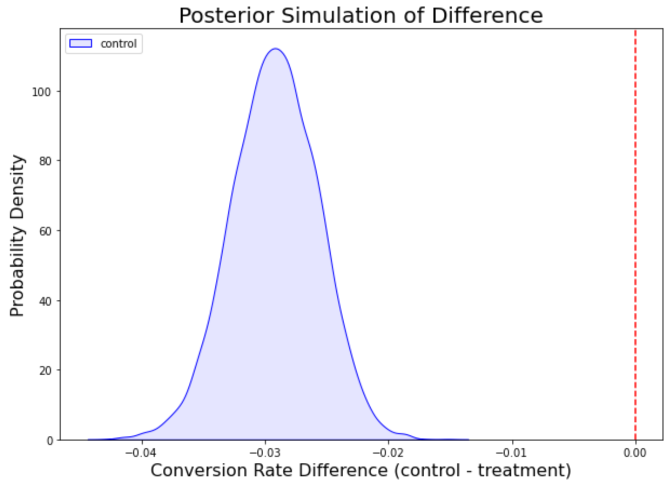
Control Posterior A vs Treatment Posterior B



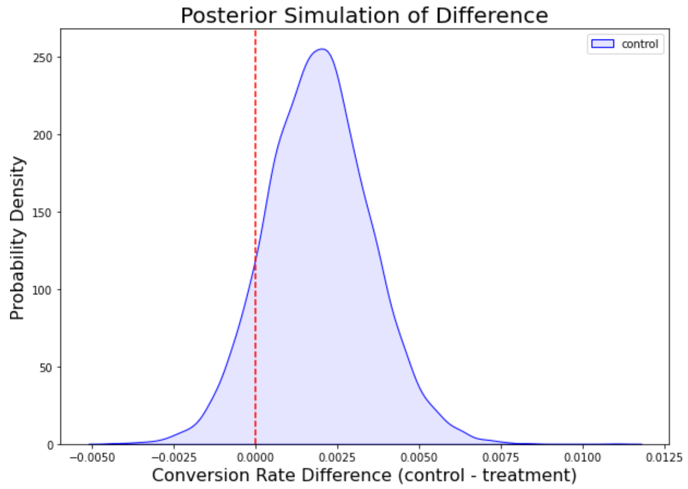
Control Posterior A vs Treatment Posterior C



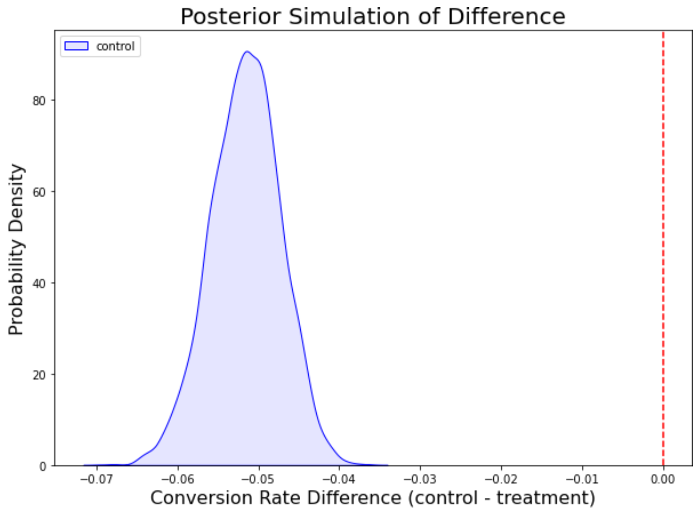
Control Posterior A vs Treatment Posterior D



Control Posterior A vs Treatment Posterior B



Control Posterior A vs Treatment Posterior C



Control Posterior A vs Treatment Posterior D

Conclusion and Solution

Conclusion and Solution

1. Conclusions

As the analysis has been done for variant group A (Control), group B (Treatment 1), group C (Treatment 2), and group D (Treatment 3), the experiment conclude that there are any significant value of CVR that capture by figure below.

	pair_group	p-value	rank	BH-crit	Significant?
5	D > C	1.489758e-35	1	0.008333	Yes
2	D > A	8.512005e-34	2	0.016667	Yes
0	B > A	4.231391e-22	3	0.025000	Yes
4	D > B	8.197395e-04	4	0.033333	Yes
1	C > A	7.668808e-01	5	0.041667	No
3	C > B	1.000000e+00	6	0.050000	No

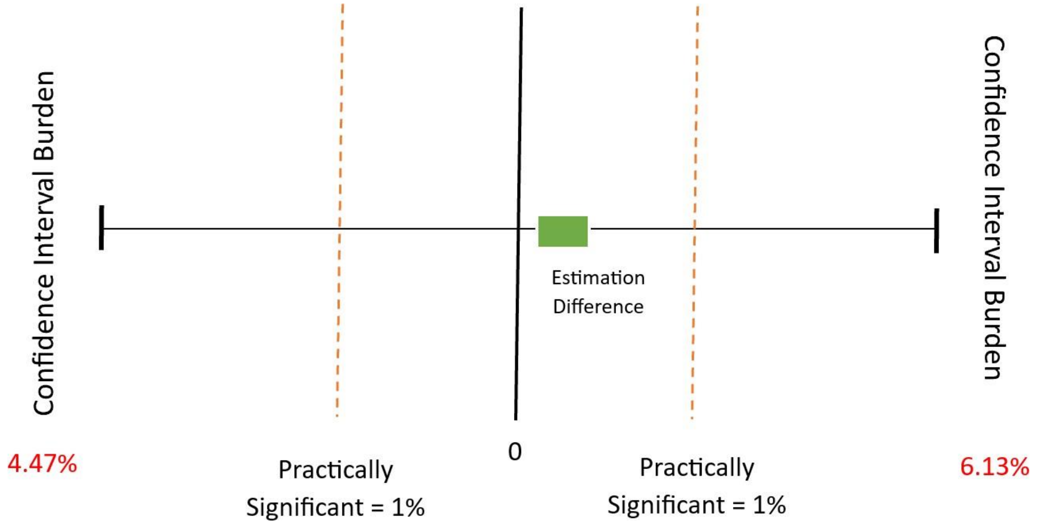
- Based on the results of multiple testing, test for group D vs group C, group D vs group A, group B vs group A, and group B vs group D resulted a significant outcome.
- Because our hypothesis is to compare whether group j is more than group i, so it can be concluded that there is sufficient evidence that the conversion rate of group D (ad + total ads > 15) is higher than groups A, B and C.
- Group D is the group that has the highest conversion rate among all groups.
- Group D becomes the winning version of the 4 combinations of the marketing company.
- It means that using creative ads with total ads > 15 statistically has an impact on increasing conversion rates.

Conclusion and Solution

2. Solution for the business (Recommendation)

The Solution that I recommend for the marketing company: based on the statistical test results, it is statistically significant. However, to make a decision whether to use creative ads or not, it needs to be ensured whether it is practically significant such as the cost of using ads, marketing costs, etc. should not incur losses.

With a minimum difference in conversion rate of 1%, the Confidence Interval values for A vs D (4.47% - 6.13%) so that the use of Creative Ads is recommended. Also, increase total ads that has been seen by customers for the



Difference between the Group A (Control) vs Group D (Treatment)

Conclusion and Solution

3. Recommendations for the future experiment

For the future experiment, there are several recommendations that can be implemented

- Define detail of Creative Ads to be variant of the experimentation such as implemented new Call to Action, etc
- Compare the engagement levels between the two groups to determine which version of the ad was more effective in generating engagement by other metrics.
- Company can try to using marketing using digital platforms to get more scalable marketing and reach broaden market that using psa or creative offline ads

References

- https://towardsdatascience.com/bayesian-a-b-testing-and-its-benefits-a7bbe5cb5103
- https://www.statology.org/benjamini-hochberg-procedure/
- https://www.indeed.com/career-advice/career-development/what-is-a-public-service-announcement
- https://mediacommons.psu.edu/2017/02/14/public-service-announcement/

Thank You