AB Testing on Marketing
Public Service Announcements (PSA) and Ads Campaign

A. Introduction & 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.

PSAs can take many forms, including television or radio commercials, print ads, billboards, social media posts, and online videos. They can address a wide range of topics, from public health and safety to environmental issues, social justice, and civic engagement.

PSAs are a valuable tool for promoting public awareness and social change because they are often produced and distributed at no cost to the organizations or individuals involved. They can also reach a large audience and have a significant impact on public opinion and behavior.

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.

The companies are interested in answering two questions:

- Would the campaign be successful?
- If the campaign was successful, how much of that success could be attributed to the ads?

With the second question in mind, we normally do an A/B test. The majority of the people will be exposed to ads (the experimental group). And a small portion of people (the control group) would instead see a Public Service Announcement (PSA) (or nothing) in the exact size and place the ad would normally be.

The idea 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

B. Setting Up Problem

- **Experimental Goal**

  The experimental goal for AB testing on marketing PSA 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 this, the experiment could be designed to compare two versions of the same advertising campaign (A and B), with one version featuring a PSA 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.

- **Choosing Metrics**

  To formulating the experimental metrics, there are several step that should be done. First, we should define the objective experiment. After that, identify how to achieve the objective. From the objective an how to achieve, we can decide the intended outcomes that want to be fulfill. After that, we can decide the Driver Metrics and Guardrail 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

  From those aspects that have been told before, the writer determines all those metrics that needed to this experiment. Here it is the outcome of the 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.

- **Define 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 Started Now", or "Buy Now"

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

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

- **Define 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

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

C. Designing Experiments

- **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.
  b. Power level (1 - β)
     1 – β = 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 - β value. Conservatively, industry rules using 80% for power lever. We're determined to use 1 - β = 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.
  e. 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.**

Then, we can determine power of the experiment by reform the formula,

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

to

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

From those equation, we can get beta = 0.22 and power in 1 week 78.19%.

Second way, we can calculate it by using `power.TTestIndPower.power()`. With this method we can get power in 1 week as 78.16%

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.

- **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

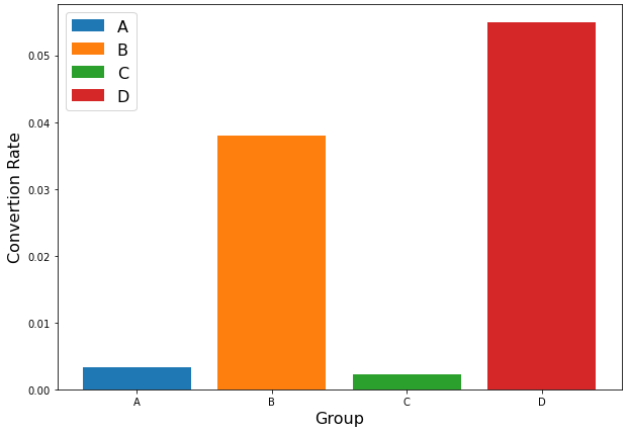
D. Running Experiment and Obtaining Data

Since we can't collect 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.

	Unnamed: 0	user id	test group	converted	total ads	most ads day	most ads hour
0	0	1069124	ad	False	130	Monday	20
1	1	1119715	ad	False	93	Tuesday	22
2	2	1144181	ad	False	21	Tuesday	18
3	3	1435133	ad	False	355	Tuesday	10
4	4	1015700	ad	False	276	Friday	14

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



	user id	test group	converted	total ads	most ads day	most ads hour	total ads group	group
549699	922187	psa	False	2	Tuesday	9	<15	A
218284	909961	psa	False	9	Thursday	14	<15	A
439360	920939	psa	False	1	Saturday	20	<15	A
527476	900626	psa	False	4	Friday	15	<15	A
500435	915897	psa	False	8	Sunday	22	<15	A
...	...	...	...	...	...	...	...	...
82046	1253451	ad	False	64	Sunday	21	>15	D
256349	1620702	ad	True	83	Tuesday	10	>15	D
569898	1376836	ad	False	24	Saturday	10	>15	D
140092	1213024	ad	True	24	Tuesday	15	>15	D
142995	1017209	ad	True	129	Friday	15	>15	D

12320 rows × 8 columns

E. Analyzing and Interpreting the Data

1. Ensure the trustworthiness

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 :

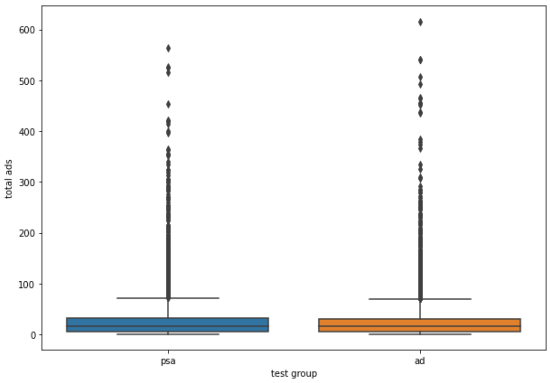
- Missing rates : How much missing value in dataset
- Uniqueness : No duplicate data
- Invalid values : Do the values follow the proper format? Are the values valid for the variable/column?

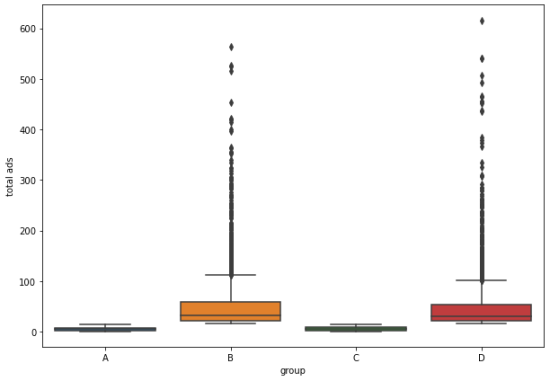
- Data delays : How many data is there at the periode of the experiment? How long does it take between when the events were logged and when the data is available for analysis?
  We're also do the checking for any NaN data (Missing value), Duplicate Data, and Invalid Data by it own combination

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

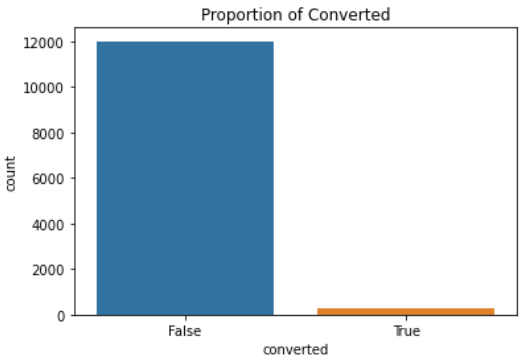
Group	# user	Percentage
A	3080	25.00%
B	3080	25.00%
C	3080	25.00%
D	3080	25.00%

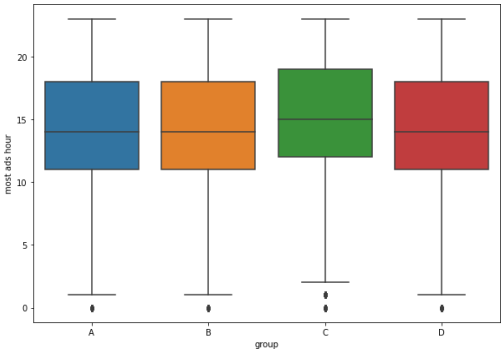
   - Using box plot, we can analysis there are no difference between total ads that consume by customer with psa and creative ads



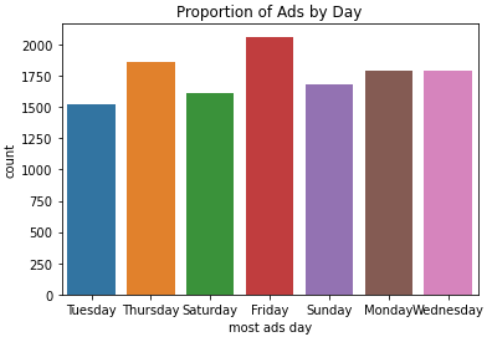


   - There are disparities of proportion between converted user in marketing funnel. And there are not significant difference between hour of ad between all variants

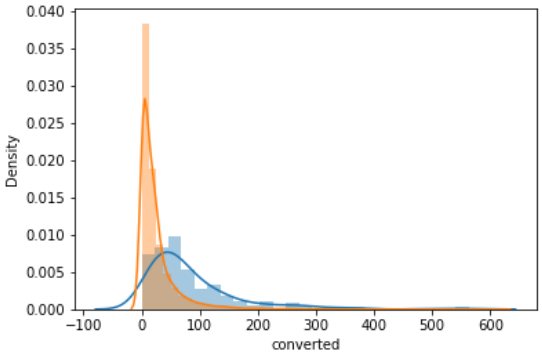


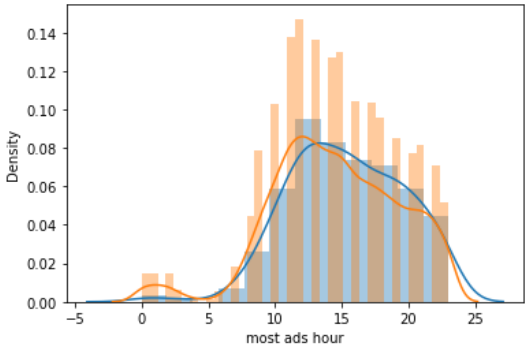


- Ads and psa has much implemented at Friday, and has less implementation at Tuesday ada Saturday.



- 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





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

Then we can calculate the chi-square statistics using the function in the `scipy` library, namely `chisquare` with steps:

- Import library
  - from scipy.stats import chisquare
- Use the function `chisquare(f_obs, f_exp=...)`
  - `f_obs`: Observed frequencies in each category (array)
  - `f_exp`: Expected frequencies in each category. By default the categories are assumed to be equally likely.

3. Define decision rules

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:

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

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.

2. Hypothesis testing and analyze the result

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-H ochberg Correction. To find out which one is the best, we can do a hypothesis testing. A suit able hypothesis test for this case is the z-test for proportion. Because we have more than tw o groups to compare, therefore we perform multiple hypothesis testing for each group pair. An issue with multiple hypothesis testing is increasing of Type I error, so we can do correctio n with Benjamini-Hochberg Correction.

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

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

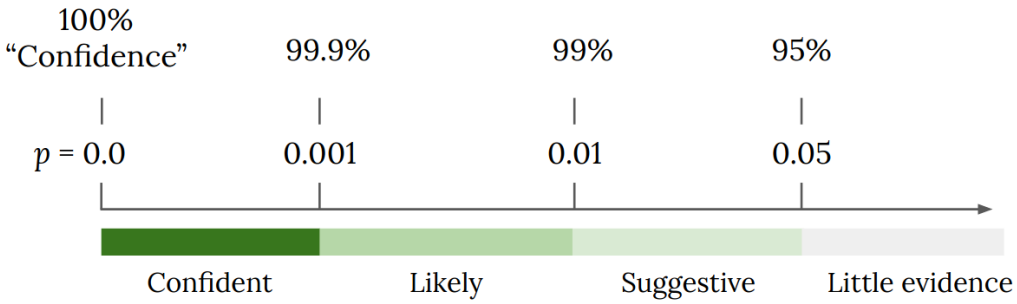
- group C vs group D

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

And then set siginificance level (alpha) = 0.05

b. Calculate the p-value in each test

Remember the rule of thumbs



Using "statsmodels.stats.proportion", we can calculate the p value of group proportions z test and generate the outcome.

	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

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

The hypothesis for each group pair is as follows:

	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

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

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.

3. Confidence 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.

4. **Calculate the Probability to Be Best (PBB)**

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).

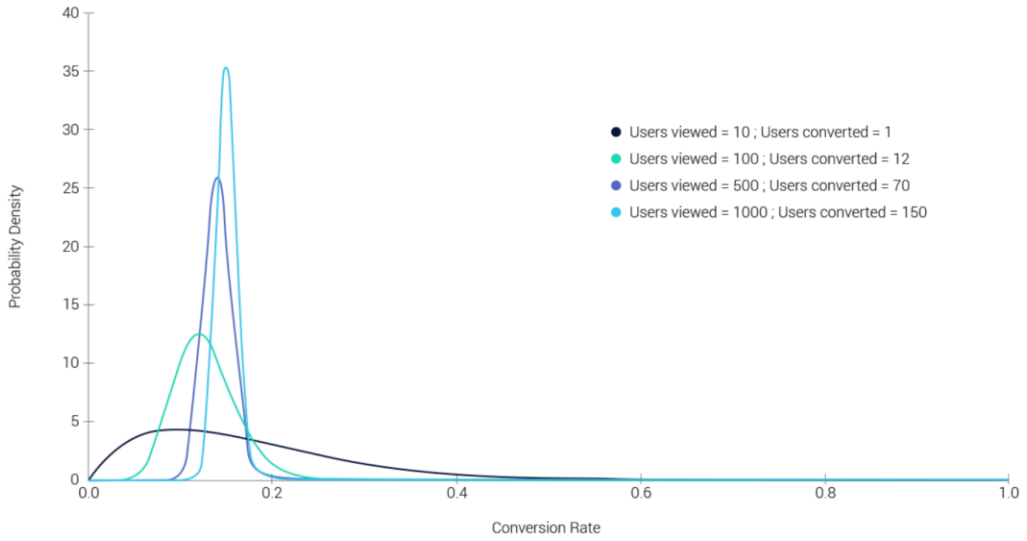
$$P(\mu|z) \propto P(\mu)P(z|\mu)$$

$P(\mu)$ 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

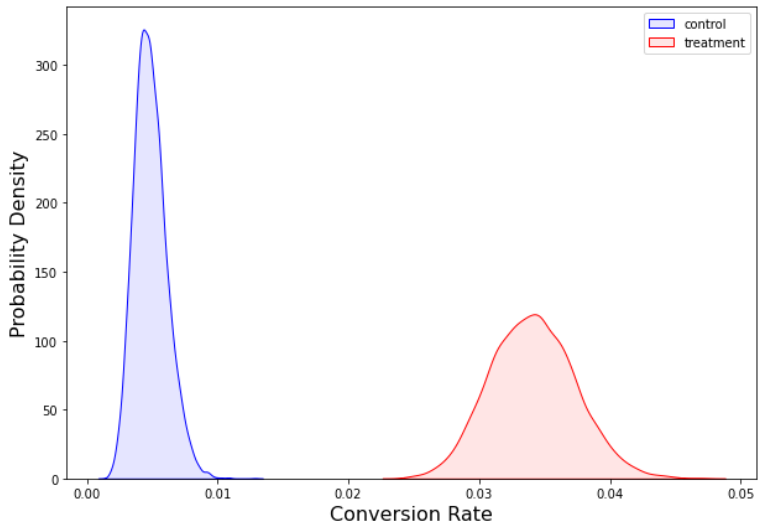


We can model the above distribution using Beta distribution. Why? Because beta distribution return value between 0-1.

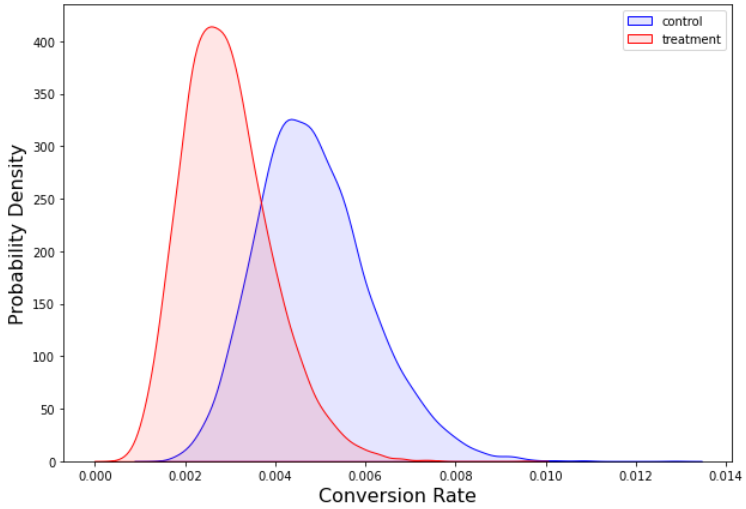
$$P(z|\mu) \sim Beta(\alpha|\beta)$$

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

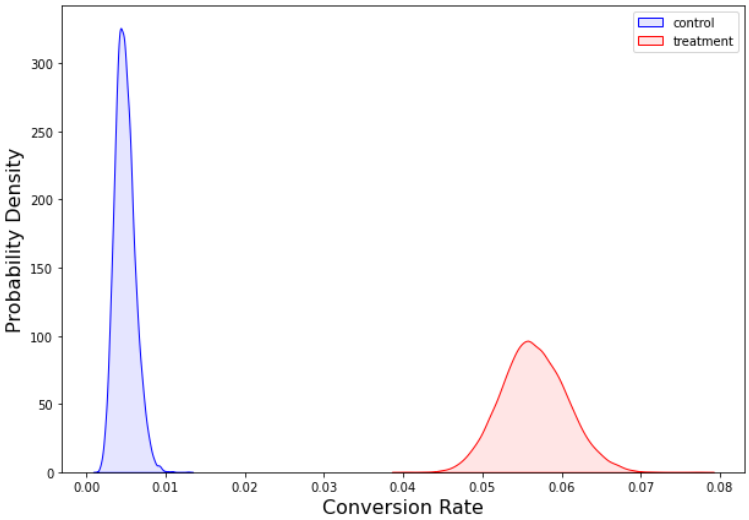
$$P(\mu|z) \sim Beta(\alpha = n_{success} + 1, \beta = n_{fail} + 1)$$



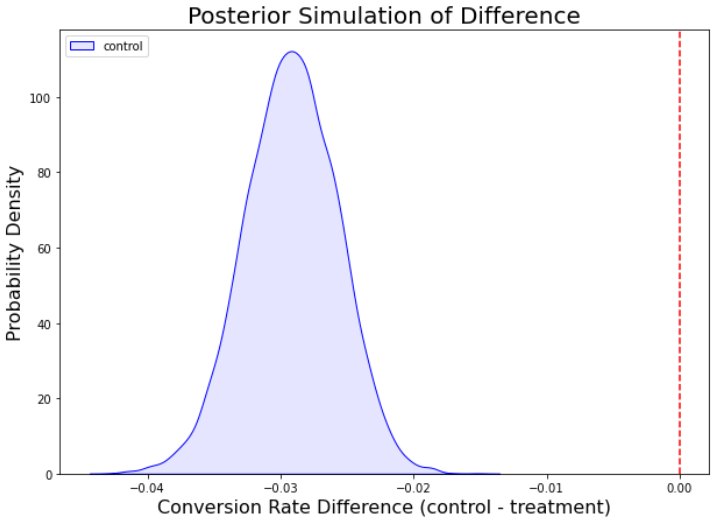
Control Posterior A vs Treatment Posterior B



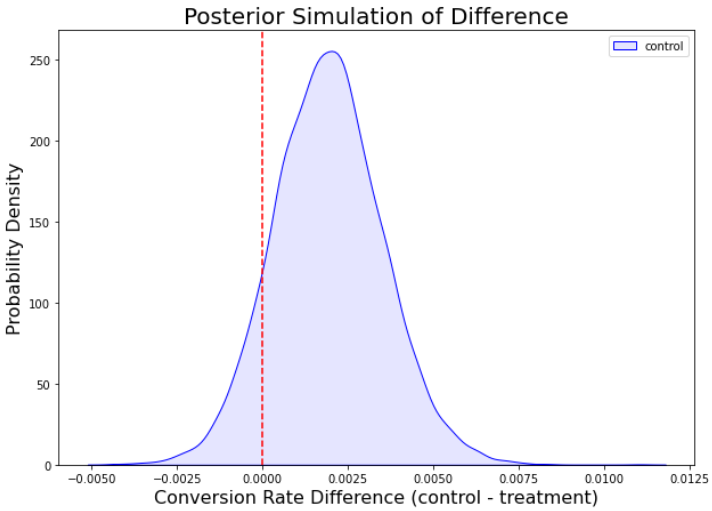
Control Posterior A vs Treatment Posterior C



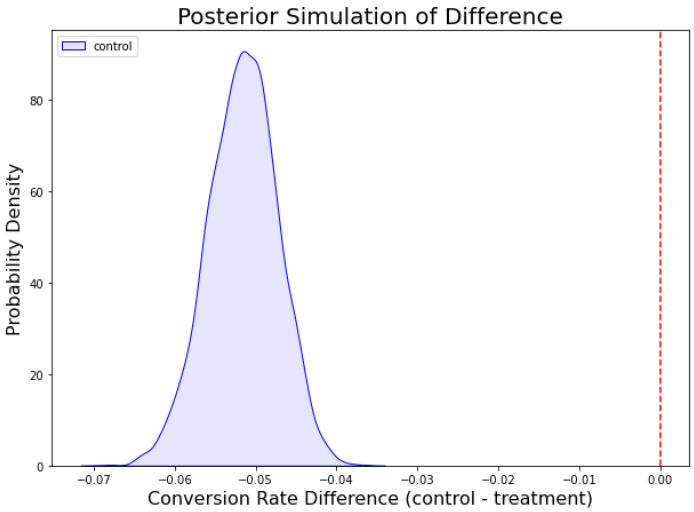
Control Posterior A vs Treatment Posterior D



Control Posterior A vs Treatment Posterior B



Control Posterior A vs Treatment Posterior C



Control Posterior A vs Treatment Posterior D

F. Conclusion and Recommendations

1. The conclusions from the previous analysis

- 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.

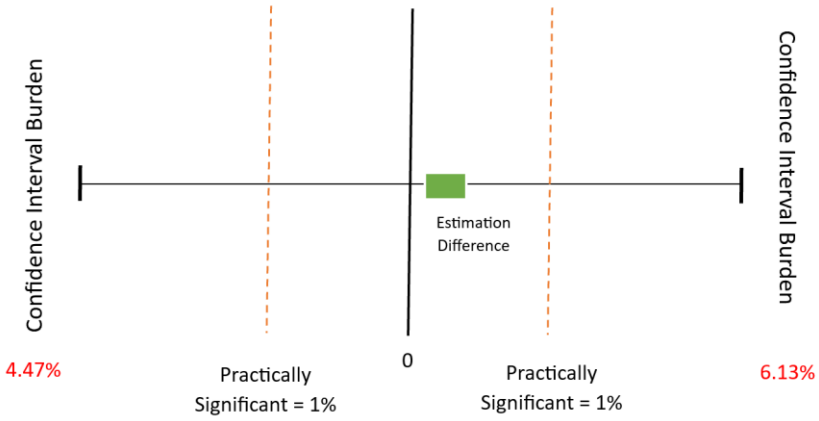
	pair_group	p-value	rank	BH-crit	Significant?
5	D > C	1.489758e-35	1	0.008333	Yes
2	D > A	8.512005e-34	2	0.016667	Yes
0	B > A	4.231391e-22	3	0.025000	Yes
4	D > B	8.197395e-04	4	0.033333	Yes
1	C > A	7.668808e-01	5	0.041667	No
3	C > B	1.000000e+00	6	0.050000	No

- 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.

2. Recommendations for the business
   Recommendation 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)

3. Recommendation for the next experiment
   For the next 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

G. Reference

- https://www.statisticshowto.com/benjamini-hochberg-procedure/
- https://blog.croct.com/post/bayesian-ab-testing
- https://www.storyboardthat.com/articles/e/public-service-announcements