A/B testing for SmartAd brand awareness

10Academy

Name: Nabil Seid
Date Aug 2020

Statistical intuition on A/B testing (pp 2-11)

Control and exposed groups

In A/B testing there are two groups. Exposed groups and control groups.

Exposed groups are test samples that are exposed to changes in independent variables being tested. The impact on dependent variables then recorded. Here users who have been shown an online interactive ad, with the SmartAd brand are in exposed groups.

Control groups are test samples that are exposed to constant independent variables being tested so that the independent variables cannot influence the results. The users who have been shown a dummy ad are in control groups.

How are the users labeled as exposed or control groups?

Depending on the context of the testing, many mechanisms can be used to label users. A very appropriate way would be to access users' cookies and label users based on the history found on the cookie. A more expensive and reliable way is to label users and store their credentials on the back-end. By compromising or assuming some variables will not affect the testing we can group users by location, browser type and handset type then label them as exposed or control groups.

Using count to make a judgement on experiment

Using count as a measure is possible when the sample data is a complete representative of the population and bothe the control and exposed groups are on the same scale (size). But that is merely ideal. There are a lot of parameters one has to account when making a judgement on an experiment like sample randomness. Thus a more sophisticated scientific computing that considers the parameters should be used to make judgement on experiment.

Statistical process to generates the data and statistical model to simulate the data

Binomial distribution is the right process to generate the data since it satisfy all 4 requirement of binomial distribution.

- The experiment consists of n identical trials.
- Each trial results in one of the two outcomes, called success and failure.
- The probability of success, denoted p, remains the same from trial to trial.
- The n trials are independent.

Logistic regression is a good model for simulating this data because it returns the fitted data as a probability brand awareness occurrence.

Appropriate statistical tests to test binomial distribution

For a large sample size we have a normal distribution of binomial probability distribution. Z-test is used when the variance is known and sample size large.

P-value, type-I error and type-II error

The p-value is the measure of the strength of evidence in support of a null hypothesis.

Type-I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**

Type-II error occurs when the researcher fails to reject a null hypothesis that is false.

P-value is related to **type-I** error and p-value is type-I error rate

classical A/B testing framework

1. Make our Hypothesis:
   - H0: there's no difference in brand awareness between the 2 groups
   - H1:There's a difference in brand awareness.
2. Set significant level (alpha)
3. Sample grouped into control and exposed groups randomly with equal probability.
4. Perform hypothesis testing to acquire p-value
5. Rejecting or failing to reject the null hypothesis. Rejection happens when p value is less than α (level of significance) else fails to reject the null hypothesis.

Sequential A/B testing workflow

In sequential sampling instead of a fixed sample size you choose one item (or a few) at a time, and then test your hypothesis. You can either:

- Reject the null hypothesis (H0) in favor of the alternate hypothesis (H1) and stop
- Keep the null hypothesis and stop
- Unable to reach either conclusion with current observation and continue sampling

Advantages of sequential A/B testing

- Optimize necessary observation(sample size)
- Reduce the likelihood of error
- Finish experiment earlier without increasing the possibility of false results

Machine learning A/B testing

- First data collectioned and preprocessed.
- Data is splitted into train and test sets.
- Fit the model using the train set and predict using the test set.
- Evaluation the model where the predicted results are matched with the actual results.
- Obtain p-values of the predictor variables from the model. The p values indicate how significant a predictor feature is towards the target variable.
- Determine whether the user group feature is significant in predicting conversion rate

Unlike statistical inference, Machine Learning algorithms enable us to model complex systems and provide a direction and magnitude of the experiment.

Classic A/B testing analysis of SmartAd brand awareness

Business Objective

SmartAd has implemented a new advertising campaign and has collectecting BIO (brand impact optimiser) data from 3-10 July 2020. The data contains “Yes” or “No” response of online users to the following question

Q: Do you know the brand SmartAd?

- O Yes
- O No

SmartAd wanted to know if the new advertising campaign resulted in a significant lift in brand awareness.

BIO Data overview analysis

For testing purposes the data was gathered as:

> ***Control***: users who have been shown a dummy ad
> ***Exposed:*** users who have been shown a creative, an online interactive ad, with the SmartAd brand.

The data is available for download here.

The data has the following variables

- **auction_id:** the unique id of the online user who has been presented the BIO. In standard terminologies this is called an impression id. The user may see the BIO questionnaire but choose not to respond. In that case both the yes and no columns are zero.
- **experiment**: which group the user belongs to - control or exposed.
- **date**: the date in YYYY-MM-DD format
- **hour**: the hour of the day in HH format.
- **device_make**: the name of the type of device the user has e.g. Samsung
- **platform_os:** the id of the OS the user has. **browser**: the name of the browser the user uses to see the BIO questionnaire.
- **yes**: 1 if the user chooses the “Yes” radio button for the BIO questionnaire.
- **no**: 1 if the user chooses the “No” radio button for the BIO questionnaire.

The data has a total of 8077 observations. Out of the total observations 6834 have not responded to the questionnaire "*Do you know the brand SmartAd*". They will not be useful for the classic A/B testing, thus we will drop them. Now we are left with 1243 observations with 586 control groups and 657 exposed groups.

Here is a summary pivot table for control and exposed groups.

experiment	aware	not aware	total	rate
control	264	322	586	0.450512
exposed	308	349	657	0.468798

Table. summary pivot table for control and exposed groups

There is a 1.8% difference in conversion rates between the two groups.

Hypothesis Testing

Now, we will conduct a hypothesis test to verify if the new advertising campaign resulted in a significant lift in brand awareness.

We will follow 5 steps to conduct a hypothesis testing.

Step 1. We define a Null hypothesis and an Alternative hypothesis.

Null hypothesis (H0): the new advertising campaign **did not result** in a significant lift in brand awareness.
Alternative hypothesis (H1): the new advertising campaign **resulted** in a significant lift in brand awareness

Step 2. We define a significant level.

We will take 5% level of significance, In most testing 5% significance level is taken to preset that there should be at least 5% measure of evidence to not reject the Null hypothesis .

Step 3. Take sample and perform statistic measures

We have a sample of 1243 observations with 586 control groups and 657 exposed groups.

From the above summary pivot table, control groups have a conversion rate of *0.451* and exposed groups have conversion rate of *0.469*

Step 4. Compute P-value

P-value is a probability value of getting the sample in step 3 given the Null hypothesis is true or the new advertising campaign did not result in a significant lift in brand awareness.

Or simply we can express it as the probability of getting a 0.469 mean conversion rate that is 0.18 away from the mean conversion rate of the Null hypothesis.

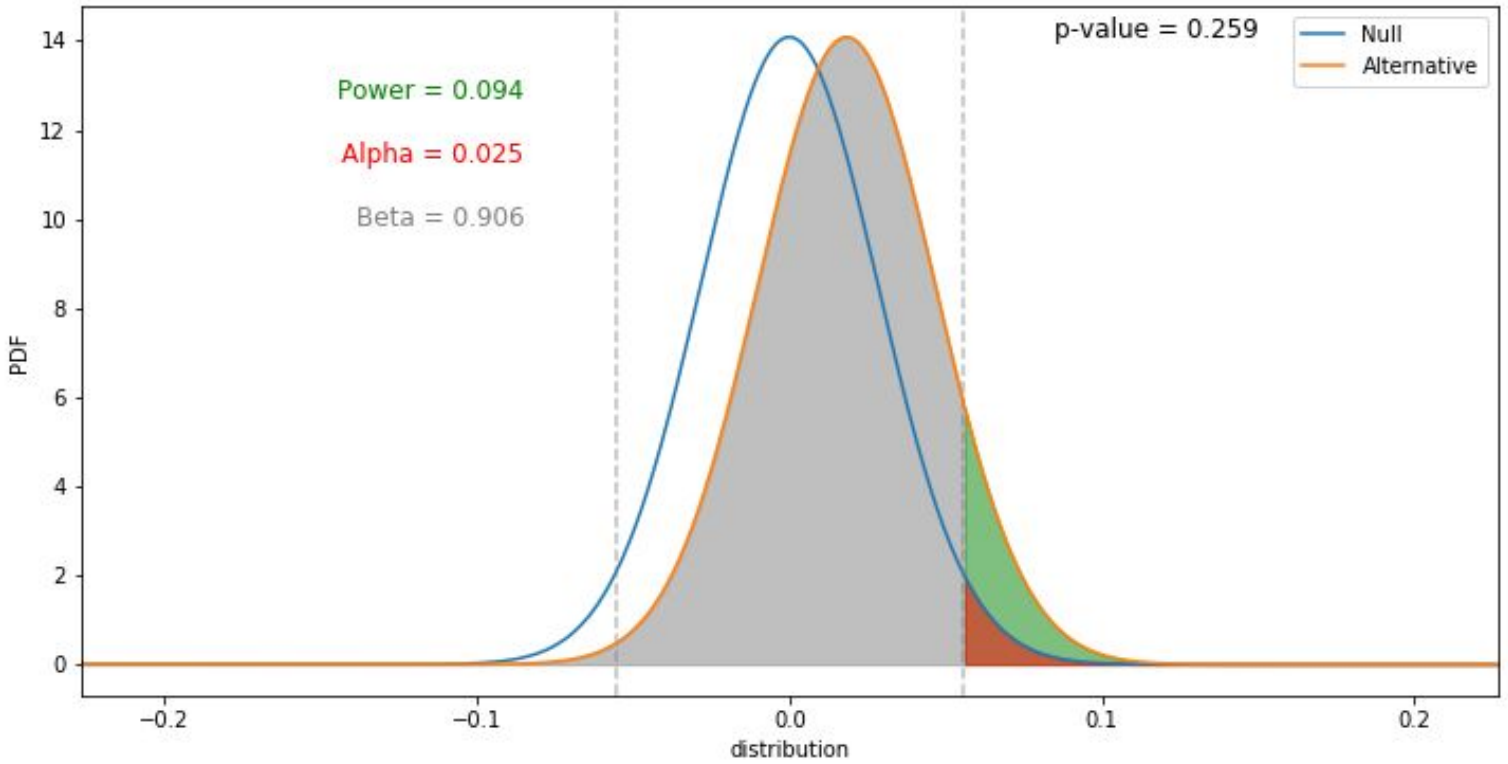
After computing the p-value turned out to be 0.259 or 25.9%, which is a very high value for a p-value. This is telling us there is a 25.9% chance of getting the sample from step 3.
25.9% is a big enough probability that we actually can get the sample by chance, thus we take this as strong evidence.

Step 5. Make a decision on the Null hypothesis based on P-value

We have a 0.05 level of significance and p-value of 0.259. P-value is greater than the significant level, thus we fail to reject the Null hypothesis. Therefore, from the experiment our result would be the new advertising campaign did not result in a significant lift in brand awareness. But we should not make a decision solely on p-value. We should consider the power value.

Power value is the probability of rejecting H0 when it is false. In simple terms it is the probability of not making type-II errors. Since it is a probability of not making errors it needs to be high enough to conduct an experiment. But we need to be careful to not make it very high. As power increases, significance level(alpha) also increases. Alpha is the probability of making type-I errors, we do not want it to be large.

The distribution of the null hypothesis and the alternative hypothesis is plotted below with p-value, power, alpha and beta values.



In the above figure we get:

Power value of 0.094 or 9.4% chance of rejecting H0 when it is actually false.
Beta value of 0.906 or 90.6% chance of failed to reject H0 when it is actually false
Alpha value of 0.025 which is the point where the left interval cross the x-axis

Having a power of 9.4% is basically telling us 90.6% of the time we will make type-II errors. Now having this information we should not make a decision from this sample. We need to have a higher power value. To increase the power we will have to gather more data. And we need to know exactly how many more samples we need to get a certain value of power.

What is optimal sample size?

Most people consider power to be 80% or 90% (just as we generally use 95% as the confidence level for confidence interval estimates).

let *ap* be average probability of both group and *cd* be conversion rate difference
we can calculate sample size with
Given significant level(*α*), beta(*β*), power(1 -*β*), detectable effect *d* and baseline conversion rate *p,* minimum sample size is

$$n = \frac{(Z_{1-\frac{\alpha}{2}} sd_1 + Z_{1-\beta} sd_2)^2}{d^2}$$

Using the above formula we get 11661 min sample size for 80% power and 15610 min sample size for 90% power.

Conclusion

Using the above sample data for making decisions, we would be wrong for 90.6% of a time. We should not use this experiment to make business decisions. We should collect more data to increase our decision success. The number of days it takes to collect a certain sample size can be roughly calculated from the sample data. Using the duration we can calculate the approximated resource it would take. Having this information we have to optimise our power value according to the available resource and conduct the experiment again.

Reference

Classic Testing

https://cosmiccoding.com.au/tutorials/ab_tests

https://towardsdatascience.com/the-math-behind-a-b-testing-with-example-code-part-1-of-2-7be752e1d06f

Plotting distribution

ttps://emredjan.github.io/blog/2017/07/19/plotting-distributions/?source=post_page-----c5ebaafdeedd----------

sample size optimization

http://statweb.stanford.edu/~susan/courses/s141/hopower.pdf

https://wise1.cgu.edu/power/power_sample.asp