A/B Hypothesis Testing: Ad campaign performance

Understanding A/B testing framework

SmartAd is a mobile first advertiser agency. It designs intuitive touch-enabled advertising. It provides brands with an automated advertising experience via machine learning and creative excellence. SmartAd provides an additional service called Brand Impact Optimiser (BIO), a lightweight questionnaire, served with every campaign to determine the impact of the creative, the ad they design, on various upper funnel metrics, including memorability and brand sentiment.

1.Which online users belong to the control and exposed groups?

The users that were presented with the questionnaire above were chosen according to the following rule:

*Control*: users who have been shown a dummy ad

*Exposed:* users who have been shown a creative (ad) that was designed by SmartAd for the client.

2.How are the users targeted?

Their company is based on the principle of voluntary participation which is proven to increase brand engagement and memorability 10 x more than static alternatives.

3.Could we use the counts of yes and no answers to make a judgment on which experiment is performing better? For example if #yes > #no for the exposed group than the control group, could we declare that the ad had a significant impact Why or why not?

To evaluate the significance of a set of experiment hypotheses, we must perform a statistical computation, such as a p-value calculation, to determine whether or not our hypothesis is significant. However, determining the significance solely on the basis of yes or no cumulative values is insufficient. Furthermore, assuming merely yes or no counted values is insufficient, especially if we are anticipating real-world difficulties. Other considerations could include whether or not users are using the app on a mobile device, which browsers they are using, and so on.

# 

4. The statistical method used is a randomized experiment.In randomized studies, chance plays a role in assigning subjects to treatment groups from a random sample. Unless the group assignments are random, no reasonable inference may be made.

5.In order to determine the significance of a set of experiment hypotheses to be true we need to do a statistical calculation like using p-value calculation to determine if our hypothesis is significant or not. But given the yes or no cumulative values to determine the significance isn’t sufficient.

In addition, assuming the yes or no counted values only, especially if we are assuming real world problems, isn't enough. There might be other factors like users using the app on a mobile or not, users using different browsers and many other factors. Based on the amount of yes or no answers isn’t sufficient to conclude.

6

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true)..In classical (frequentist) A/B testing, we use p-values to measure the significance of the experimental feature (being exposed to an ad in our case) over the null hypothesis (the hypothesis that there is no difference in brand awareness between the exposed and control groups in the current case).

A P-value is composed of three parts:

1.The probability random chance would result in the observation.

2.The probability of observing something else that is equally rare.

3.The probability of observing something rarer or more extreme.

The p-value is the sum of the three probabilities.

A small P-value will tell us to reject the null hypothesis. Typically we only reject a hypothesis if the P-value is less than 0.05.

How do you know if a p-value is statistically significant?

The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

However, if the p-value is below your threshold of significance (typically p < 0.05), you can reject the null hypothesis, but this does not mean that there is a 95% probability that the alternative hypothesis is true. The p-value is conditional upon the null hypothesis being true, but is unrelated to the truth or falsity of the alternative hypothesis.

A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis. You should note that you cannot accept the null hypothesis, we can only reject the null or fail to reject it.

A statistically significant result cannot prove that a research hypothesis is correct (as this implies 100% certainty).

The probability of making a type I error is represented by your alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis.You can reduce your risk of committing a type I error by using a lower value for p. For example, a p-value of 0.01 would mean there is a 1% chance of committing a Type I error.

7.How does the classical A/B testing (using z-test, f-test, etc.) framework work?

A/B tests are often used to experiment with page design options that vary dramatically, including the position of text and pictures, background colors, number of pictures on the page, use of icons, and navigation structure. The tests usually involve fewer combinations with more extreme changes.The classic A/B test presents users with two variations of your pages at the same URL. That way, you can compare two or several variations of the same element.Randomly subset the users and show one set of the control and one the treatment. Monitor the conversion rates of each group to see which is better.

8.Sequential A/B Testing Workflow

* Start your experiment with choosing sample size, let’s call it N;
* Randomly assign variations under test to the treatment and control, with 50% probability each.
* Track the number of incoming successes for both variations. Let’s refer to the conversion rate of treatment variation as T, and CR of control as C.
* It’s necessary to finish the test when T−C reaches √2N and declare the treatment variation to be the winner of your A/B test.
* It’s necessary to finish the test when T+C reaches N. In such a case, declare that the experiment had no winner.

9.

Sequential A/B testing might become a robust alternative. Such experiments don’t only optimize necessary traffic volumes but also reduce the likelihood of mistakes

when it comes to classic A/B testing, it’s allowed to check test results only at the very end when the sample size for both variations is reached. Sequential A/B testing in its turn allows multiple checks on every step ensuring that the error level won’t exceed 5%.

10.

11.A major issue with traditional, statistical-inference approaches to A/B Testing is that it only compares 2 variables - an experiment/control to an outcome. The problem is that customer behavior is vastly more complex than this. Customers take different paths, spend different amounts of time on the site, come from different backgrounds (age, gender, interests), and more. This is where Machine Learning excels - generating insights from complex systems. Unlike statistical inference, Machine Learning algorithms enable us to model complex systems that include all of the ongoing events, user features, and more. There are a number of algorithms each with strengths and weaknesses.

An attractive benefit to Machine Learning is that we can combine multiple approaches to gain insights.

Con: It May Take Time (and Resources) for Machine Learning to Bring Results.In other words, machine learning takes time, especially if you have limited computing power. Handling tremendous volumes of data and running computer models sucks up a lot of computing power, which can potentially be quite costly. So, before turning to machine learning, it’s important to consider whether you can invest the amount of time and/or money required to develop the technology to a point where it will be useful.

12.