DCONJOINT

A Mu Sigma University Publication, December 2014

Mu Sigma

In This Issue...

Abstract

A/B testing has evolved as the proven method to take effective decisions in online business. But what is the Math which explains the accuracy of the test results? Validity of the test results depends on the statistical parameters and the expected values defined for them. As online businesses are reaching end users through multiple channels, adaptation of A/B testing methodology has also undergone significant technological evolution.

Basics of A/B Testing

- By Bhagavath R.C.

Introduction

The definition of A/B testing can be stated as the controlled experimentation of change(s) in the web page's design, architecture, structure or content and comparing it with current design, architecture, structure or content to determine and understand positive or (maybe) negative impact. It can be said to be something as simple as the old picture game of finding six differences between two images which are almost similar but with a few subtle differences. A/B testing is a real world scenario of math and technology driving business decisions. Let us explain how.

The concept of A/B testing is nothing new, it has been adopted from the "Test and Control" treatments in the pharma industry - in which a new medicine or vaccine will be given to a test group of patients and a placebo to another control group and the recovery rate is monitored at the same time. The new medicine will be accepted as a cure only when the recovery rate of the test group is significantly higher or different from that of the control group.

Similarly, any web interface which monetizes or determines conversions based on user visits, clicks, call-to-action or sign-ups in the site, has to employ A/B testing to understand whether any decision of change in the web site will positively drive the site's conversion rate. Let's consider an example - e-commerce site ABC.com has been thinking to move its "Buy Now" button from the bottom right corner of the site to the bottom left corner as the change has been made by a competitor and leading e-tailer, XYZ.com. The executives have decided to run an A/B test to figure whether they should be making a change to their site. But before that they should be answering the following questions:

- What is the baseline? For comparing the results of the A/B test, current baseline has to be accurately **defined and determined**. For ABC.com, it'll be the number of users who are currently clicking the "Buy Now" button in the right corner of the site.
- What is the 'test metric'? This is crucial for the success of any A/B test scenario. For ABC.com, after making the change in the site, the increase in conversion rate of the overall site may be negligible compared to the conversion rate of users moving from products page to the "Place Order" page. The correct interpretation and intended impact of the test results depends on the 'test metric' for which the A/B test is conducted. At the same time, one should always be keen to identify any kind of ripple effect on the overall performance metrics too.
- How confident are you with the A/B testing results? To re-phrase the question, will there be same results when the change is made on 100% traffic? Or will it be different? If yes, what should be an acceptable difference?

DCONJOINT

A Mu Sigma University Publication, December 2014

Mu Sigma

In This Issue...

These are the very questions which this article intends to answer. The credibility of the test results directly depends on the statistics of the sample, population and method of sampling.

Sampling and Sample Size

Sample size is the amount of traffic in which the changed website (wherein the "Buy Now" button will have been shifted to the corner) will be shown. For an A/B testing scenario, most recommended traffic size for test and control is 50-50. But it may not be a wise thing to convert web-page for 50% percent traffic blindly. So, the website has to come up with an ideal percentage of traffic in which the test can be conducted. Will it be 2% or 20%? Again that depends on the population. For a site which has 100 visitors, considering 2% traffic as test means only 2 visitors will get to see the new page. This can lead to drastic conclusions.

Without the ideal sample size, the chances of committing a Type 1 or Type 2 error* is more. The base sample size can be estimated by defining the following parameters:

- What is the current baseline or current value of test metric? (current conversion rate)
- What is the expected value of test metric? (expected conversion rate)
- What should be the significance level (1% or 5%) and statistical power?

All statistical tools provide functions to come up with solutions for these problems. In R, for example, you can use the **power.prop.test** function to calculate power, sample size or significance level.

The next thing which one needs to be careful about is sampling. The samples compared should ideally differ only with respect to the webpages they are visiting. This may be virtually impossible as users may use different machines or delete cookies or may not login to the site while visiting, all creating sample bias. Still there can be simple measures which can be taken to reduce this sample bias.

- Time of testing: Both test and control has to be run in the same time and also for the same time period. Comparing test and control metrics executed at different points of time and for different time periods will not yield the correct results due to the inherent bias of seasonality associated with different times. It has to be made sure that test and control variants are available to the users in the same time such that they are received by almost similar number of visitors.
- Randomness of the sample: The sample is identified based on cookies, login information or machine information (IP/machine id). Even if these info have chances of being changed or altered over time, random sampling should be employed to make sure that the sample is a sample not biased by location, hardware and software features.

Confidence Intervals and Statistical Significance

After conducting an A/B testing experiment, the success and failure depends on the results; how different are the results from each other and is the difference statistically significant.

DCONJOINT

A Mu Sigma University Publication, December 2014

Mu Sigma

In This Issue...

The average value and the confidence interval of the metric for test and control scenarios will be obtained and actual difference will be calculated. Confidence interval is the margin of error calculated for both set of test and control metrics. The expectation is that when the change is applied, 100% traffic will be within the confidence levels. When there is an overlap of confidence intervals of test and control metrics it can be either because the sample sizes are not large enough or change made is not significant to make an impact. If the confidence intervals are not overlapping even if there is a high difference between the average metrics of test and control, it doesn't mean it's significant. It has to be either compared with the current baseline for relative difference or one has to use the t-test to identify whether the two samples are different.

What's Next in A/B Testing

Multivariate testing is the experimentation of several page attributes in the same time. It can be compared to conducting multiple A/B testing experiments in one go. For example, in the above case, where we talk of relocating the "Buy Now" button for ABC.com, if the background has also to be changed from white to either red, blue or green, then, one will need to do 2(for buttons)*3(BG colors) = 6 experiments along with the control experiment to identify the best combination of attributes for the site. As multiple attributes are being tested, multivariate testing is more complex and time consuming than A/B testing. More data has to be collected to validate the results for multivariate testing. Still the technical implementation for both are relatively same.

A/B Testing in Mobile

The world has moved from desktops to laptops to mobiles, so has the idea of A/B testing. Mobile App developers have been utilizing A/B testing tools for mobile apps by several providers for some time now. A/B testing is conducted to see the change in conversion rate by changing the color of the button for online shopping apps to check the ideal speed of the race car in a car game app. Some of the providers have options for multivariate testing too. As mentioned above, since validation of multivariate testing requires huge amounts of data, multivariate testing is recommended for apps with a large user base. Major technology players like Facebook and Amazon are providing tools for Mobile App A/B testing.

The process for setting up A/B testing in mobile apps is more or less similar across operating systems. The app developer or publisher has to integrate the software development kit (SDK) with his/her app and push the updated version to the respective app store. The app developer can test different features of the app using the interface provided by the tool provider. If a particular change has to be tested, the update is pushed to the target sample (developer, who can select a particular target sample based on segments created on app usage or feature usage or game level).

DCONJOINT

A Mu Sigma University Publication, December 2014

Mu Sigma

In This Issue...

This feature is not available with all testing tool providers. Once the app update has been installed by the relevant user, the interface will start getting data and validation for effect of the change that can be done after collecting the required amount of data. Major difference of A/B testing in Mobile and Web-pages is that the former cannot be real time in some cases. If an app update with an A/B test change is pushed to the app store, there may be a delay as the app store has to approve the update. And the data will be received only if the user has downloaded and installed the update. These two scenarios may increase the entire testing time period. While for webpages, any changes integrated will be experienced by the user from his or her next visit to the site.

Conclusion

A/B testing has been and continues to be one of the most effective ways of testing new features and altering existing features, hence driving business decisions for webpages and mobile apps. For conducting any successful A/B test, following parameters play a very vital role: feature to be tested, metrics to be measured, sample size, sampling, significance level and third party app testing tool provider (for mobile apps). Failure of a test doesn't mean that the site can't be improved further. Also testing shouldn't be stopped after one significant output, it may fail for the next sample size. Testing should be a continuous process to make sure the site is improving and also increasing the user experience, in which the business owner has to decide the levels of improvements by choosing the right parameters all the time.

Notes

Type 1 error : Incorrect rejection of null hypothesis which is actually true is called type 1 error or error of first kind.
Type 2 error : Accepting null hypothesis which is actually false is called type 2 error or error of second kind
http://en.wikipedia.org/wiki/Type_I_and_type_II_errors