A/B Testing: Challenges with Statistical Analysis

Johan Settlin and Joar Ekelund

***Abstract*—Continuous development is important for websites and other software products to stay relevant and keeping customers satisfied. One of the great challenges with introducing new features, services or a simple style change is knowing the perceived change in value for the end users. A/B Testing offers a way to test these changes, and give a basis for measuring the perceived change in value. A/B Testing can also be used to find versions of the software that increase sales, or increase time spent with the software. This is done by gathering user data for different versions of the software and measuring how the change is perceived according to some predefined metric. A statistical analysis is then performed to see if the change had a positive effect.**

This literary study will explain what A/B testing is and the logic behind it, as well as the benefits and challenges associated with A/B testing. A major focus will be on the statistical challenges and how to overcome them. The conclusion reached is that A/B testing offers a powerful and systematic way of measuring customer satisfaction, user engagement and conversion rates which can be used to minimize the risks associated with taking design decisions. Although A/B Testing has a lot to offer it also exist a lot of challenges that needs to be addressed before it can be used, example of challenges include knowing how much data needs to be collected, for how long to run the experiment, and what elements to change between version to be able to perform a meaningful statistical analysis.

I. INTRODUCTION

One of the risks with software development in highly dynamic situations is that the created product offers users little to no value, meaning that the time spent developing the product was wasted.

Companies need ways to evaluate the customer value in their products, one way of doing this is to continuously execute experiments and collect customer input and data as a part of the development process [1]. To understand how to create costumer value or what it means for a large number of customers is not trivial in practice. Creating customer value with a competitive set of great product features is not always enough. Companies needs to understand their customers' needs as well as behavior and create tailored solutions for their costumers needs. [2].

Marjo Kauppinen et. al describes in the paper ”From Feature Development to Customer Value Creation” [2] that there is a few pitfalls when trying to create customer value and also some practices to create costumer value. Some of the pitfalls mentioned in the paper is:

- Focusing on launching features as fast as possible
- Adding too many features making the product more complex, this could even decrease the costumer value.
- Creating products that does not support the costumer process, the products are not tailored to the need of the users.

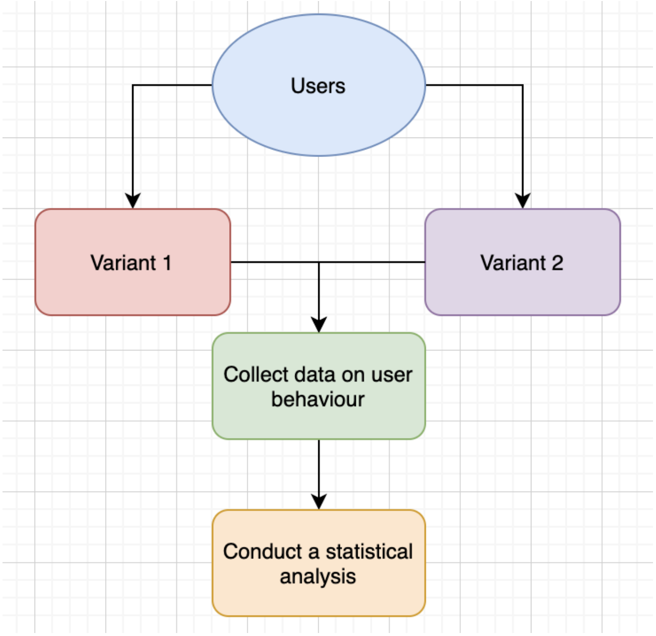
Users
Variant 1
Variant 2
Collect data on user behaviour
Conduct a statistical analysis

Fig. 1: An basic overview of how A/B testing is done. Figure created by the Authors.

In addition some of the practises mention to create costumer value is:

- Discover information about customer processes actively.
- Identify customer segments. The basic idea of customer segmentation is to analyze existing and potential customers of the product. Based on the analysis, the product can be customized for each customer segment.

A/B Testing is an experiment driven development approach that could reduce the risk of having low costumer value. This is done by providing developers with data from users to continually improve the offered features and services or when introducing new ones. A/B Testing removes the guesswork and allow developers to make data driven decisions based on user information.

All the decisions made are based on a statistical analysis which is the backbone of A/B tests. When conducting a statistical analysis there are challenges that needs to be taken into account to assure a valid result. The objective of this essay is to introduce A/B testing to the reader, take an in-depth look at statistical analysis and the challenges that statistical analysis brings up. The reader will get an understanding of how A/B testing can be done and which pitfalls to avoid when conducting a statistical analysis on their A/B tests.

Benefit	Motivation
User engagement	A/B Testing can help finding a version of the software that is more engaging to users, this can be done by for example measuring clicks generated by users or time spent on the different versions of the software [5].
Higher conversion rate	One important metric for any company is the conversion rate, meaning that visitors to for example a website end up as customers. With the help of A/B testing different versions of the software can be tested and conversion measurements can be taken, and more effective implementations can be made [5], [6].
Risk reduction	Making a change in software could lead to a lower perceived value for end users. A/B testing allows for different alternatives to be tested and evaluated. This removes the guesswork and minimizes the risks associated with implementing new features or services [5].

TABLE I: A summary of the benefits with A/B Testing

II. A/B Testing

A/B Testing, also known as bucket testing, split-run testing or controlled experiment; is a method used to evaluate user engagement or satisfaction [3]. The method is widely used by several different companies such as Netflix and Facebook [3].

A/B Testing is performed by randomly splitting users of a service into groups. Each group is presented with similar versions of the software with the exception of a key element of interest. While the users from different groups engage with the software data is collected. After the data is collected a statistical analysis is performed and the different versions are compared against some key metric to see which version performed the best. An overview of the different steps in A/B testing can be seen in fig. 1.

A simple example of this would be a website with two different versions. The difference between the versions could be the style of a button. The users of the website are presented with one of the two versions and information about whether the button is clicked or not is collected. A statistical analysis is performed to see which of the two designs were preferred going by the engagement shown from the users by clicking the button.

This simple example is very similar to one that was used for President Obama's election campaign, for the 2008 election. For Obama's campaign website several different versions of a button were used alongside different videos and images. A picture of Obama with his family and a button with the text “Learn More” was shown to improve the sign up for campaign information with 40.6 percent when compared to the original website. This simple change translated to an estimated 57 million dollars extra in donations for his campaign. [4]

A. Benefits

By conducting A/B testing there can be a lot of benefits under the condition that the experiments are executed properly. You can get more customer satisfaction, sell more product, more clicks etc. Some of the potential benefits by performing A/B testing to websites [5], [6] are listed in table. I. In table. I benefits listed are user engagement, higher conversions rate and risk reduction. The first two benefits are related to customer satisfaction and the last benefit describes the advantage of removing guesswork in business decisions.

It is important to know that these benefits is not a guarantee and that A/B testing isn't some magical thing that always lead to success, the effect of the test may show that the current variant performs better and thereby all that came out of the experiment was spending money on something that ended up not being used. Therefore, conducting these kinds of experiments continuously may be a good way to reduce this risk since you learn from the user data and can design better tests in the future and make data driven decisions.

B. Challenges

Even though A/B testing have a lot of benefits there are some challenges in doing controlled experiments on websites with users' data. To achieve the best results the challenges of A/B testing needs to be understood to maximize the results of the experiment.

The aspect of whether the company have the capabilities or not to implement several version at once and to be able to iteratively make improvements on the products based on the test results [7]. This can be a costly and time-consuming process especially if it is done continuously at the company.

There is also the aspect of the technical difficulties and challenges in conducting a good and reliable experiment.

Challenge	Motivation
The amount of collected data	In the common case the collected data has a high variance which means that early results can often be misleading Thereby enough data needs to be collected to get good measurements [9].
The duration of the experiment	Some experiments may benefit from being active for a longer period of time and the effect may be that the confidence interval shrinks and thereby increasing the statistical power (eg. click through). However, in some cases this is not true, and the duration is not very relevant (eg. sessions per user) and the amount of users is more important [8].
Keep the design simple	If the experiment is too complex there is many pitfalls to run in to. There can be hidden bugs and normally the complexity is not necessary. Keeping the experiment simpler makes the results more trustworthy [9].
Performance	The performance matters a lot, if one of the candidates in the experiment has slower response time than the other candidate this will have a high impact in the result [9].

TABLE II: A summary of the challenges with A/B Testing

Kohavi et al. have discussed challenges and issues with A/B testing in their research [8], [9]. Some of the challenges mentioned includes data collection, duration of experiments, the design of the experiment and performance of the test. A summary of their results can be seen in table II. Important factors are planning and execution of the experiment and how to handle the collected data to get the best results. Another big part of A/B testing is statistical analysis, an in-depth discussion about statistical challenges and solutions can be found later in the essay.

III. Statistical Analysis

As mentioned before A/B testing divides users of a service into groups, collect data, perform a statistical analysis and make a data driven decision for which version is the best. This section contains information about the statistical aspects of A/B testing. This include Null Hypothesis Statistical Testing, distributions and the main challenges associated with statistical analysis in A/B Testing.

A. Null Hypothesis Statistical Testing

The statistical analysis used for A/B Testing can be performed in several different ways, but one of the most common is Null Hypothesis Statistical Testing (NHST) [10]. NHST is an approach for deciding between two interpretations of a statistical relationship. The two interpretations are known as the null hypothesis H_0 and the alternative hypothesis H_1. The null hypothesis is often that there is no statistical relationship in the population, and that if such a relation is seen in the sample it is due to random chance, while the alternative hypothesis often is that there exists such a relationship. After the hypothesis have been decided the likelihood of the sample given that H_0 is true is calculated. If the sample relationship is highly unlikely the null hypothesis is rejected in favour of the alternative hypothesis. The likelihood of the sample, given the null hypothesis is called the p-value. A low p-value indicate that the sample is very unlikely given H_0. Before a test is conducted and the p-values calculated it is important to define an alpha value, also known as a significance level. This value represents how low the p-value must be for us to reject H_0 in favor of H_1. If this occurs the result is said to be statistically significant. If on the other hand the p-value is greater than the predefined alpha we've “failed to reject the null hypothesis”, note that this is not the same as “accepting the null hypothesis”. [11]

B. Distributions

Depending on what data the A/B test examines different statistical distributions can be used. S. Borodovsky and S. Rosset describes in their paper ”A/B Testing at SweetIM: The Importance of Proper Statistical Analysis” [12] that when the A/B test for example examine binary outcomes such as click or no click on a button, a binomial distribution [13] is often used for statistical inference and calculations. Sometimes the data in the A/B test follows a continuous trait, this could for example be time spent on a website, in that case a normal distribution [14] is often suitable to use. If the test measure counts such as number of searches by a specific user then the data often follows a Poisson distribution [15].

After the data is collected a NHST as mentioned before is conducted to determine whether you should keep alternative A or if the hypothesis is rejected and thereby consider using alternative B. However, there are some challenges in conduction a statistical analysis.

C. Statistical Challenges And Solutions

In the article ”Top 5 mistakes with statistics in A/B testing” [16] Georgi Georgiev describes some of the statistical challenges associated with A/B Testing.

One common challenge is to attribute a business goal, such as increasing sales, to a certain metric. One could easily think that this would be as easy as measuring the clicks of a buy button between the different versions. However, using this as a measurement could actually lead to a version which decreases sales. Imagine that we're comparing the current version of our website with a proposed new version and we can prove that the proposed version generates significantly more clicks than the current one. We therefor conclude that the new version is better. What we failed to realize is that the new version, while generating more clicks, has a much lower average value of each individual order. We end up with more orders, but lower profit since each order is less valuable. A more suitable metric would be to use the number of clicks multiplied with the average order value. This metric better represents the business goal since the generated conclusion would actually factor in the differences in profit. [16]

Another common statistical mistake made is using two-tailed tests when a one-tailed test is sufficient. Recall that when performing NHST we reject the null hypothesis H_0 if the the p-value is sufficiently small. If we take a simple example of measuring clicks between a current and a new version, it becomes clear that the p-value would be low if the new version generates a lot of clicks or very few. Since we're interested in improving the product we often don't care to differentiate between if the new version is worse or equally as good as the current. This is what is called a one-tailed test, when we only care about one part of the spectrum and a two-tailed test is of course when were interested in both ends of the spectrum. There are two main problems when using a two-sided test when a one-sided test would suffice. The first problem is that a two-sided test requires more data which means that the experiment has to run for longer before a conclusion can be reached. The second problem is that it overestimates the uncertainty meaning that we would not reject the null hypothesis even though the desired significance level is achieved, see fig. 2 for an illustration of this. [16]

When performing A/B Testing we want to use statistics to draw conclusions about all users of the service from a sample. To do this the sample needs to be representative of the population. In the article ”14 AB Testing Sample Size Issues and Mistakes That Can Ruin Your Test” [17] written by Khalid Saleh several examples of data pollution are mentioned that could invalidate the results of the tests. There are several reasons why the test data could become polluted, some are even outside the testers control. One common mistake is to run too short tests, this could potentially lead to a biased sample due to promotions, ad campaigns or pay day of customers. On the other hand, if a test runs for too long this could cause problems with the partitioning of users. For a web-based example these partitioning errors could be caused due to deletion of cookies, resulting in testers not being able to differentiate between new or old users, or that a user uses several devices and is therefore presented the different versions of the website. This makes it hard to determine which version was preferred by the user. Saleh recommends a test period of 4 weeks to avoid the issues with too short tests and the problems with cookie deletion due to too long ones. This however is a rule of thumb and not an all-purpose solution. [17]

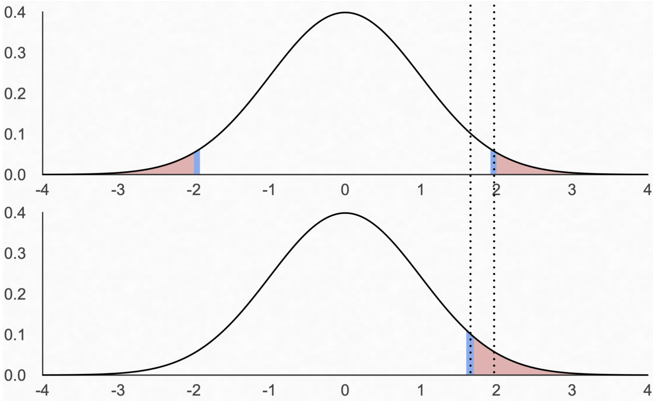

Fig. 2: Two-tailed test on top. One-tailed test on bottom. Both are normal distribution with mean 0, standard deviation 1 and alpha value 0.05. Leftmost dotted line indicate the boundary where we reject H_0 for the one-tailed test, the second dotted line represent the same boundary for the two-tailed test. Figure created by the Authors.

Saleh also mentions that recurrent users could be biased and prefer the old version even though the new version is better. It is therefore preferable to use new users for the tests. This requires that data about recurrent users are available and that the service has enough traffic to use only newcomers for the experiment. [17]

IV. Conclusion

As with most things, A/B Testing comes with both benefits and drawbacks. When used correctly it can be an extremely useful tool improving customer experience, sales, interaction or time spent using the software. The possibility to make data driven decision removes a lot of guesswork and gives developer a better understanding of users' needs as well as a foundation to compare implementations in an objective manner. As shown from the Obama campaign, simple changes can have large real-world impacts and finding these small tweaks would be hard without the metrics provided by A/B testing. Statistical analysis plays a great part in A/B tests and enables us to achieve all the potential benefits.

On the other hand, A/B testing can be both costly and ineffective if the right metrics are not used or if the experiments are not carefully planned. To achieve statistically significant results a lot of challenges exist, some are even beyond the control of the tester. There are several factors to take into consideration and external data pollution from a competitors ad campaign can't be helped. To give the best chances of non-polluted data a test should be performed over a period such that unavoidable interference can be neglected while still being able to control that a user sees the same version for each interaction. A resistance to change is also an important aspect, old users may prefer a worse implementation simply because it is familiar and one should try, if possible, to base tests on newcomers.

A/B testing seems to be a great way to generate potential customer value and to lower the risks for company's decision making on both design and content for websites and other software products. In an agile environment or within DevOps teams where the focus is to shorten the software development cycle and to continuously deliver high quality software, A/B testing is a great practice to reduce the risk of implementing unnecessary features or worsen the current version of the software. If A/B testing is used continuously in the development process metrics created from customer data can be used to reduce these risk when continuously deliver software if the statistical analysis is well performed.

References

[1] S. G. Yaman, F. Fagerholm, M. Munezero, J. Münch, M. Aaltola, C. Palmu, and T. Männistö, "Transitioning towards continuous experimentation in a large software product and service development organisation – a case study," in *Product-Focused Software Process Improvement* (P. Abrahamsson, A. Jedlitschka, A. Nguyen Duc, M. Felderer, S. Amasaki, and T. Mikkonen, eds.), (Cham), pp. 344–359, Springer International Publishing, 2016.

[2] M. Kauppinen, J. Savolainen, L. Lehtola, M. Komssi, H. Tohonen, and A. Davis, "From feature development to customer value creation," in *2009 17th IEEE International Requirements Engineering Conference*, pp. 275–280, 2009.

[3] Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. Bhasin, "From infrastructure to culture: A/b testing challenges in large scale social networks," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, (New York, NY, USA), p. 2227–2236, Association for Computing Machinery, 2015.

[4] D. Siroker and P. Koomen, *A/B testing: The most powerful way to turn clicks into customers.* John Wiley & Sons, 2013.

[5] BrightEdge, "What is a/b testing?," *online: https://www.brightedge.com/glossary/benefits-recommendations-ab-testing*, 2020.

[6] L. Kolowich, "How to do a/b testing: A checklist you'll want to bookmark," *online: https://blog.hubspot.com/marketing/how-to-do-a-b-testing*.

[7] P. Hynninen and M. Kauppinen, "A/b testing: A promising tool for customer value evaluation," in *2014 IEEE 1st International Workshop on Requirements Engineering and Testing (RET)*, pp. 16–17, 2014.

[8] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu, "Trustworthy online controlled experiments: Five puzzling outcomes explained," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, (New York, NY, USA), p. 786–794, Association for Computing Machinery, 2012.

[9] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu, "Seven rules of thumb for web site experimenters," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, (New York, NY, USA), p. 1857–1866, Association for Computing Machinery, 2014.

[10] A. Deng, J. Lu, and S. Chen, "Continuous monitoring of a/b tests without pain: Optional stopping in bayesian testing," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 243–252, 2016.

[11] P. C. Price, I.-C. A. Chiang, R. Jhangiani, *et al.*, "Research methods in psychology: 2nd canadian edition," ch. 13, pp. 249–253, 2018.

[12] S. Borodovsky and S. Rosset, "A/b testing at sweetim: The importance of proper statistical analysis," in *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 733–740, 2011.

[13] Wikipedia contributors, "Binomial distribution — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Binomial_distribution&oldid=952672118, 2020. [Online; accessed 16-May-2020].

[14] Wikipedia contributors, "Normal distribution — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Normal_distribution&oldid=956659144, 2020. [Online; accessed 16-May-2020].

[15] Wikipedia contributors, "Poisson distribution — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Poisson_distribution&oldid=956083134, 2020. [Online; accessed 16-May-2020].

[16] G. Georgiev, "Top 5 mistakes with statistics in a/b testing," *online: https://towardsdatascience.com/top-5-mistakes-with-statistics-in-a-b-testing-9b121ea1827c*, 2019.

[17] K. Saleh, "14 ab testing sample size issues and mistakes that can ruin your test." https://www.invespcro.com/blog/ab-testing-14-sampling-issues-that-can-ruin-your-test/, 2020.