A/B Testing: Benefits And Challenges

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 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. 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, 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 user 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 customer 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 and behaviours 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.

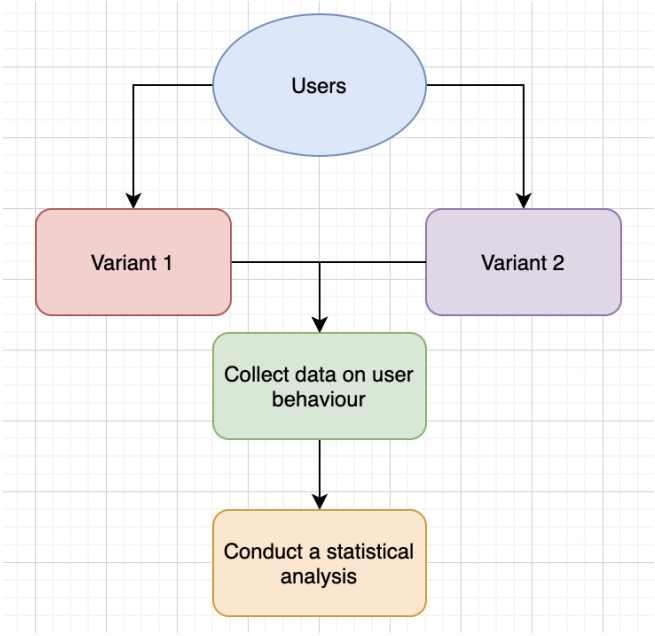

Fig. 1: An basic overview of how A/B testing is done

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.

II. A/B TESTING

A/B Testing, also known as bucket testing, split-run testing or controlled experiment, is 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 groups 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

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.
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.
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.

TABLE I: Benefits with A/B Testing

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 is presented with one of the two versions and information about whether the button is clicked is collected. A statistical analysis is performed to see which of the two designs were preferred going by the engagement show 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 was 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]

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. The statistical analysis can be performed in several different ways, but one of the most common is Null Hypothesis Statistical Testing (NHST) [5].

NHST is an approach to deciding between two interpretations of a statistical relationship. The two interpretations are known as the null hypothesis H0 and the alternative hypothesis H1. 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 exist such a relationship. After the hypothesis have been decided the likelihood of the sample given that H0 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 H0. Before a test is conducted and the p-values calculated it is important to define an alpha value. This value represents how low the p-value must be for us to reject H0 in favour of H1. 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". [6]

IV. 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 [7], [8] are listed in table. I.

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 kind 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.

V. Challenges

Even though A/B testing have a lot of benefits there are some challenges in doing controlled experiments on websites

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.
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.
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.
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.

TABLE II: Challenges with A/B Testing

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.

There is the aspect of if the company have the capabilities to implement several version at once and to being able to iteratively make improvements on the products based on the test results [9]. 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. Kohavi et al. have discussed challenges and issues with A/B testing in their research [10], [11]. A summary of their results can be seen in table II.

VI. 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.

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. One important aspect is that the experiments should not be too complex because it can be very hard to draw any conclusions from the gathered data. Another thing is that the duration of the experiments needs to be long enough to not get misleading results.

A/B testing seems to be a great way to generate potential customer value and to lower the risks for companies 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 practise to reduce the risk of some of the pitfalls mentioned in the introduction to create customer value such as launching features fast or adding unnecessary features. If A/B testing is used continuously in the process these metrics can be validated against customer data and thereby reduce the risk of ending up in these pitfalls even though software is continuously delivered.

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] 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.

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

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

[8] 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*.

[9] 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.

[10] 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.

[11] 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.