LUND UNIVERSITY

LTH

FACULTY OF ENGINEERING

ROI Calculation on Online Controlled Experiment

Degree Project in Mathematical Statistics for Engineers

FMSM01
Date: 18th of June 2019
Author: Richard Wang
Supervisor: Fredrik Olsson
Examiner: Magnus Wiktorsson

Abstract

As online services such as e-commerce and mobile applications keeps growing, the need of optimizing the user experience does as well. By conducting Online Controlled Experiments, companies can get an insight to which features, design choices and implementations that users enjoy the most. This projects explores two areas in relation to Return on Investment (ROI) calculations for Online Controlled Experiments. First, challenges and pitfalls with calculating a reliable ROI metric are described as well as references to research working to mitigate these challenges. The challenges presented are all related to how to accurately measure the effect between two candidate variants which sums up the main challenges with ROI calculation. Secondly, a model based on expected return on investment is constructed and explored in order to investigate whether the model can help to optimize test parameters for a two sample t-test in the setting of Online Controlled Experiments. The results of the model analysis shows that the model has limited practical use since it maximizes the ROI - quotient without taking into account to magnitude of potential revenue increase as well as potential cost.

Key Words: A/B testing, Online Controlled Experiments, Return on Investment, ROI

Contents

1 Introduction

1.1 Background

As the number of online services such as e-commerce and online advertising keeps growing, along with proliferation of data generated by user engagement of these services, opportunities to conduct controlled experiments to empirically test different versions of the services increases. Tests conducted usually entails but are not limited to; design variation such as color and layout changes to the inclusion or exclusion of certain features, functions or technical implementations.

Today, online controlled experiments are used to great extent across the industry to improve online services. Technology giants such as Amazon, Facebook and Google conduct more than 10.000 experiments annually and has reported large gains from the experiments [Kohavi and Thomke, 2017]. For instance, in 2008 Microsoft conducted a test where users would be redirected to a new window/tab when clicking on the Hotmail link on the MSN home page instead of staying in the same window. Initial tests confined to 900.000 UK users showed a user engagement increase measured by number of clicks made on the MSN home page by 8.9%. Later in 2010, the same test was conducted on the US market with 2.7 million users where the same metric showed a 5% increase. Till this day, this simple technique is still widely in use by major websites such as Facebook and Twitter.

By conducting systematic controlled experiments with scientific rigor, companies can aid their software implementation, user interface and user experience decisions by iteratively testing different versions. As such, decisions are made based on data rather than "Highest Paid Person's Opinion", which is shown to yield inferior results compared to controlled experiments [Kohavi et al., 2009].

1.2 Purpose, Problem Definition and Scope

Initiated by a consultancy firm active in the Conversion Rate Optimization (CRO) business, Conversionista, the inception of this project was rooted in the lack of literature in measuring return on investment (ROI) on online controlled experiments. A perceived challenge in the industry is to demonstrate and communicate the benefits of controlled experiments to decision makers or clients and to translate the benefits into business value. This thesis aims to examine the possibility to construct a ROI framework, or some other similar framework, that can help to quantify financial value in the context of online controlled experiments and to mitigate said challenge.

1.2.1 Research Questions

To formalize the objectives, the report aims to investigate the following research questions:

RQ1 Describe the challenges with ROI calculations in the context of online controlled experiments

RQ2 Propose a mathematical model for ROI calculations or other similar model with the purpose to quantify financial value of conducting tests

1.2.2 Method

In order to answer **RQ1**, a literature study will be conducted to understand current research and challenges within the industry. A study based on interviews was considered, however a literature review was deemed more suitable due to mainly two reasons. Firstly, extensive literature on the subject is available rendering other information sources unnecessary. Secondly, given that the problem and project stems from the industry, a review of academic sources is warranted.

The aim with the literature review is to provide insights as to what problems are currently being focused on and if these problems can explain the challenges with ROI calculations. In the literature review, mostly accepted peer reviewed papers will be used as a references and these sources will be considered validated and credible. To some extent, publications from industry vendors will be referenced. These sourced are deemed credible if independent third-party has validated the source, if such validation is not available, the credibility of each source will be evaluated on a case by case basis before being used.

The literature search will be performed with the help of general search engine, Google, as well as search engines specialized in academic publications, Google Scholar and LUBsearch[1].

For **RQ2**, based on the insights gained from the literature review, in conjunction with discussion with Conversionista, a suitable model will be presented and analyzed as seen fit depending on the nature of the model presented.

1.2.3 Delimitations

There is considerable amount of literature devoted to experiment design, statistical methods for measuring controlled experiments and ensuring statistical significance, as well as literature on best practices on technical infrastructure and organizational requirements for successful implementations of online controlled experiments. See for instance the recommended reading list provided in the entry on Online Controlled Experiments and A/B Testing in Encyclopedia of Machine Learning and Data Mining [Kohavi and Longbotham, 2017]. The literature review in this report will primarily focus on subjects related to modelling ROI, that is statistics and data science, although other aspects will be mentioned.

Furthermore, in the realm of online controlled experiments, depending on the desired objective of an experiment different key metrics can be measured. Common key metrics could for instance be fraction of newsletter receivers who engages in the letter content, fraction of people exposed to an ad who clicks on the ad or the fraction of visitors who end up buying a product on an e-commerce

[1] Search engine used internally by students and faculty members at Lund University

website. This thesis will solely focus on the last-mentioned metric, henceforth generally referred to as conversion rate.

2 Online Controlled Experiments

The following section briefly describes online controlled experiments along with some preliminary statistics to provide context to the reader.

2.1 Primer

Online Controlled Experiments, also called to as A/B testing, split tests, randomized experiments, control/treatment tests and online field experiments [Kohavi and Longbotham, 2017], refers to the practice of conducting controlled experiments in order to detect causality between changes made to a service or product and the response of a key metric being measured, often referred to as Overall Evaluation Criterion (OEC). In controlled experiments, all variables except the OEC is assumed to be constant between test subjects as opposed to uncontrolled experiments. The test population, i.e. users of the service[2], are randomly assigned to a control group or to a treatment group, the experiment is run, and causality is either confirmed or rejected, see Figure 1.

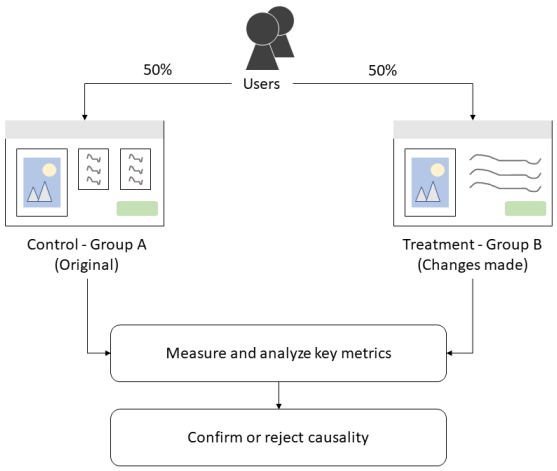

Figure 1: High level overview of online controlled experiment with one treatment group and one factor in each group.

The difference of the OEC between the control and treatment are calculated along with confidence intervals. The difference in OEC is commonly referred to

[2]Technically, users can for instance be identified by browser-stored cookies [Hohnhold et al., 2015]

as uplift in the industry and measures the relative increase (or decrease) of the OEC. In a correctly designed experiment, exogenous changes such as seasonality and competitor actions, will be evenly distributed across the groups and any statistically significant difference in the OEC between control and treatment can be inferred to have been caused by the intentional change made.

Note that more than one treatment group can be present in an experiment, this have implications on e.g. sample size to ensure statistical robustness of the tests, see section 2.2. The interested reader is referred to the seminal work "Controlled experiments on the web: survey and practical guide" [Kohavi et al., 2009] for a more comprehensive introduction to the field.

2.2 Preliminary Statistics

Let $\overline{X}_A$ denote the observed mean of the OEC, e.g. conversion rate or revenue/user, of the control group A and $\overline{X}_B$ the observed mean of OEC for treatment group B. Define the test - statistics as

$$t = \frac{\overline{X}_B - \overline{X}_A}{\widehat{\sigma}_{AB}} \tag{1}$$

Where $\widehat{\sigma}_{AB}$ denotes the empirical standard deviation of $X_B - X_A$. Since the sample size for each group are more than often in the thousands in the setting of online controlled experiments, the t-statistic as defined above converges to the Normal distribution by the Central Limit Theorem[Deng et al., 2014]. With other words, the t-test essentially becomes a z-test.

A rule of thumb formula often used to calculate required sample size is provided below for significance level $\alpha = 0.05$ and power $\beta = 0.1$ [Wheeler, 1974]

$$n = \left(\frac{4r\sigma}{\Delta}\right)^2 \tag{2}$$

Where Δ is the minimum absolute effect that one wishes to detect, $r \geq 1$ the number of factors and σ the standard deviation of the OEC. Note that in the case of conversion rates, the conversion rates follow a Bernoulli distribution[3] and as such $\sigma^2 = p * (1 - p)$. Furthermore, note that the quotient of the required sample size between two minimum detectable values is proportional to the quotient of the values squared. For instance, allowing a 5% minimum detectable effect (MDE) instead of 1% reduces the required sample size by a factor of 25 which has obvious effect on the required user traffic for conducting a test.

Following from this, confidence intervals can be calculated and hypothesis testing performed using standard classical theory [Box et al., 2005]. Since even

[3]Consider one visitor to an e-commerce website. The user can either buy something - convert, or leave the website without converting - no conversion. Let $X_i = \{1, 0\}$ where 1 and 0 indicates conversion or no conversion respectively for user i. Then, X follows a Bernoulli distribution with parameter p i.e. probability to convert. Thus, $\mu = p$ and $\sigma^2 = p * (1 - p)$.

small absolute changes often translate into non-trivial relative changes, paired with the often large traffic volumes associated with online services, it is often interesting to look at relative changes. For instance, an increase by 0.2 percentages in conversion rate from 2% to 2.2% corresponds to a 10% relative increase. In theory, ceteris paribus, this equals to a 10% revenue increase.

Defining the relative change test statistics and coefficients of variation as

$$t_{rel} = \frac{\overline{X}_B - \overline{X}_A}{\overline{X}_A} * 100\% \tag{3}$$

$$cv_A = \frac{\widehat{\sigma}_A}{\overline{X}_A} \tag{4}$$

$$cv_B = \frac{\widehat{\sigma}_B}{\overline{X}_B} \tag{5}$$

We have that the confidence interval of the relative effect [William and Briggs, 2006] is given by

$$CI_{rel} = (1 + t_{rel}) \frac{1 \pm z_{1-\alpha} * \sqrt{cv_A^2 + cv_B^2 - (z_{1-\alpha} cv_A cv_B)^2}}{1 - z_{1-\alpha} cv_A^2} - 1 \tag{6}$$

The formula assumes no covariance between the two tested groups. Also, cautious use of the formula is advisable when the confidence interval of the denominator in above formula contains zero since there is a risk that one or both endpoints doesn't exist.

2.3 A/A - tests

As the name suggests, running an experiment under the same conditions as one would in an actual A/B – test as described in section 2.1, but instead of using two different variants, uses the same variant, is called A/A – tests. There are two reasons for conducting A/A - tests [Kohavi et al., 2009]; when calculating statistical power, the variability of the OEC can be estimated through A/A - tests. Secondly, statistical assumptions as well as biases can be checked through A/A - tests. Recall that by definition, false positives (type 1 error) should occur with probability roughly equal to the set significance level α. Many experiment setups have been found to fail this litmus tests for many different reasons [Kohavi and Longbotham, 2010]. Two examples are variance underestimation due to correlated experiment units that violates assumptions for standard variance calculations, and various biases introduced by e.g. different user browsers or bots.

3 Statistical Pitfalls and Challenges when Measuring Effects

The following section describes statistical pitfalls and challenges that have been observed by practitioners in the industry along with proposed remedies from previous research. Despite many of the pitfalls being well known statistical phenomena, they are still often mentioned in the online controlled experiment literature as common mistakes to avoid.

3.1 Early Stopping and Continuous Monitoring

Early stopping also called peeking, refers to the practice of continuously monitoring a test that is performed using classical frequentist null hypothesis testing, and stopping the test the first time a p – value under the chosen significance level is observed as opposed to running the test until the pre-defined sample size is reached. The mistake is caused by practitioners wanting to minimize the cost of potentially running tests on non-performing treatments, or alternatively, wanting to declare an early winner with the intention of directing all traffic to the seemingly better variant as soon as possible in the case of observing a low early p-value.

This is of course bad practice and will inflate false positives (type I error). To quickly see why, consider two experiments on the same data; the first of which is a properly performed test where the p-value is evaluated only once when the pre-defined sample size is reached, the second where one continuously monitors and evaluates the p-value along the course of the experiment. Clearly, the second experiment have many more instances where the p-value is evaluated and thus the probability of falsely rejecting the null hypothesis is strictly increasing. It can be shown that the realized false positive probability can easily be inflated to five to ten times that of the chosen significance level when stopping early [Johari et al., 2017]. The immediate remedy is of course to simply stick to the pre-defined sample size.

The above-mentioned reasons for wanting to stop experiments early are clearly rather attractive from a business perspective and alternative frameworks have been discussed in the setting of online controlled experiments to allow for optional stopping times. Sequential tests using mixture sequential probability ratio test (mSPRT) have been discussed [Johari et al., 2017], however in a recent paper [Ju et al., 2019], it is pointed out that the framework only allows for the detection of $H_0 : \theta_0 = \theta_1, H_1 : \theta_0 \neq \theta_1$, i.e. unable to determine which variant is better. Furthermore, the same paper points out that the mSPRT has a power of one, meaning under any ground truth $\theta_0 \neq \theta_1$, the test will eventually reject the null hypothesis if the tester waits long enough. This implies that the experiment runs the risk of running a long time which is undesirable for online testing.

A/B testing using Bayesian hypothesis testing have also been discussed where the authors of the paper apply a stopping rule based on Bayes factor and controls for the false discovery rate (FDR) instead of type I error [Deng

et al., 2016a]. As always in the case of Bayesian statistics, a prior distribution is needed. Although interesting, the forever on-going discussion concerning Bayesian versus Frequentist statistics are outside the scope of this thesis. The authors do however point out, that priors can be learned objectively thanks to the availability of vast empirical data in A/B-testing.

3.2 Multiple Comparison Problem

The multiple comparison problem is a statistical artifact followed directly from type I error in a multivariate test setting. The multiplicity in an online controlled experiment can refer to various elements, for instance dividing the test population into multiple segments based on e.g. browser type [Dmitriev et al., 2017], testing for multiple treatments or multiple features, or a combination of these. For instance, consider a A/B/4 – experiment with five features, i.e. an experiment with four treatments and five different variations of the features in each treatment tested against one control group with the five features, see Figure 2. Under $\alpha = 0.05$, we expect one of the features in one of the treatments to result in a false positive assuming no correlation structure and no true difference in conversion rate.

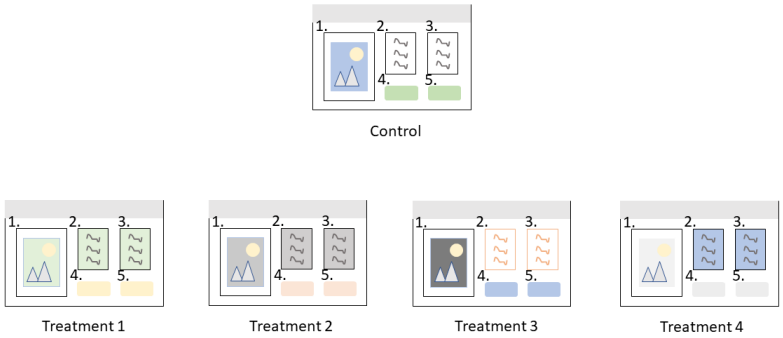

Figure 2: One control and four treatments with five variations of each features in every treatment tested against the control variations of the features. In total, 20 tests are performed.

There are two well-known approaches to control for the multiple comparison method [Johari et al., 2017]; The Bonferroni correction and the Benjamin-Hochberg procedure. The Bonferroni correction controls for the family-wise error rate (FWER), that is the probability of at least one type I error. The Benjamin-Hochberg procedure controls for the false discovery rate (FDR), that

is the expected proportion of incorrectly rejected null hypotheses. Controlling for FWER is known to be more conservative than controlling for FDR.

In the context of online controlled experiments, a method combining Bonferroni correction and Fisher's method, which combines p-values from several independent hypothesis test and combines them to one test statistics, as well as using known distribution properties have been proposed [Deng et al., 2014]. According to the author of the aforementioned article, the method is strictly less conservative than Bonferroni correction. For detecting heterogeneous treatment effects and controlling for multiple comparison problem, two methods controlling for FDR has been proposed under the assumption that the true model follows a linear regression with Gaussian errors [Xie et al., 2018].

3.3 Simpson's Paradox

Consider the results from following hypothetical experiment:

	Cumulative Conversion Rate
Control	2.90%
Treatment	2.72%

Table 1: Results at the end of a hypothetical experiment

It appears that control resulted in better conversion rates than treatment. Consider the following table segmented by days from the same hypothetical experiment:

	M	Tu	W	Th	F	Sa	Su	Tot
Conv. (C)	33.6	31.4	29.1	26.8	26.0	12.7	9.0	168.7
Vis. (C)	990	980	950	900	800	700	500	5820
C.r. (C)	3.39%	3.21%	3.07%	2.98%	3.25%	1.82%	1.79%	2.90%
Conv. (T)	0.38	0.68	1.65	3.11	7.0	7.2	12.1	32.1
Vis. (T)	10	20	50	100	200	300	500	1180
C.r. (T)	3.80%	3.42%	3.30%	3.11%	3.50%	2.40%	2.42%	2.72%

Table 2: Visitors (Vis.) is increased over time in Treatment (T) and subsequently traffic is decreased over time in Control (C). Table shows number of conversions (Conv.) and visitors in thousands along with conversion rate (C.r.).

The treatment performed better than control during all days, despite having a lower cumulative conversion rate. This statistical phenomenon is known as Simpson's paradox and it is a rather known concept among statisticians and data scientists. The above simplified example illustrates a method used in practice where the allotted traffic to treatment is increased over time when there is an uncertainty of the treatment potentially causing large negative effects. This way, the large negative effects can be discovered early without impacting user experience across large portions of the users [Kohavi and Longbotham, 2010].

Some other examples [Crook et al., 2009] where Simpson's paradox can occur are:

- Non - uniform sampling, e.g. users from certain browser are sampled at a higher rate
- Treatment and control allocation vary across segments, for instance across countries
- Only a fraction of the top spenders are allotted to treatment due to concerns of negative impact to the top spenders

In general, one should be cautious when the proportion between segments are different between control and treatment or when segment proportions changes over time. Other instances of Simpson's paradox is for instance when local metrics shows bad performance but the site-wide impact might be positive [Xu et al., 2015].

3.4 Novelty and Primacy Effects

Novelty and and primacy effects refers to the phenomena when actual effects takes time to materialize. This can arise mainly due to two reasons; one is due to the novelty effect when users sees new features and explore them out of curiosity rather than any actual value add to the user. Conversely, some features takes time for the users to learn before showing up as any positive change, thus primacy effects refers to some kind of learning effect [Dmitriev et al., 2017]. There have been some model proposals to account for this [Hohnhold et al., 2015] [Lu and Liu, 2014]

3.4.1 Non-existent Novelty and Primacy Effects

Sometimes, the data can falsely be interpreted as novelty or primacy effect. Since test subjects arrives sequentially in an online setting, the sample size will grow over the course of the experiment. Naturally, the confidence interval will decrease over time as the number of subjects increases for most metrics. Thus, perceived trends can materialize even under ground truth $\theta_0 = \theta_1$. Figure (3) illustrates the phenomenon through simulation.

The danger of the non-existent novelty and primacy effects is to extrapolate perceived trends naively and prematurely, e.g. with simple regression. It has been reported that even practitioners proficient in statistics have been fooled by this effect [Kohavi et al., 2012].

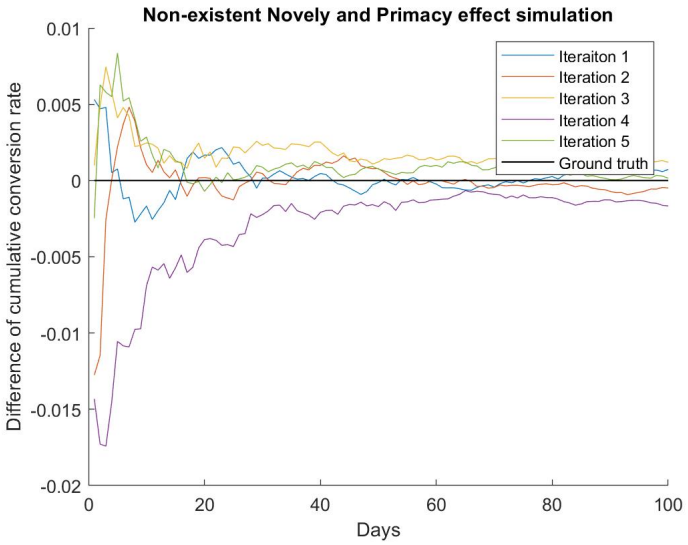

Figure 3: The graph shows five iterations of the difference of the cumulative conversion rate between two groups where visitors in each group is generated daily from i.i.d N(1000, 50) and rounded to the nearest integer value and number of conversions generated from i.i.d. Bin(Users generated by normal distribution, 0.05) during the course of 100 days. The none-zero convergence bias comes from the rounding error.

3.5 Non - decreasing Confidence Intervals

Assuming i.i.d subjects, same sample size and variance across control and treatment, we expect the confidence interval to be proportional to roughly $\frac{1}{\sqrt{n}}$, and as such get narrower confidence interval the longer the experiment. However, for some metrics, e.g. percent changes for session/user [Kohavi et al., 2012] we rarely see large declines for the confidence interval. The reason for this is that in cases such as this, the confidence interval will roughly be proportional $\frac{CV}{\sqrt{n}}$ [4] where both the mean and the standard deviation is increasing over time. The resulting effect will be a confidence interval that does not decrease over time. For most metrics, especially bounded metrics such as clickthrough rate or conversion rate, this is not a problem. However, for some count metrics as in the given example, the only way to reduce variance is to increase the sample size in control and treatment.

3.6 Variance reduction and metric sensitivity

Reducing uncertainty of the metrics in an online controlled experiment is of course desirable as in any other statistical setting. Reducing the variance and

[4] Coefficient of variation as defined in equation 4

increasing metric sensitivity implies that statistical power can be reached with smaller sample sizes and thus allowing for faster experimentation.

One of the most influential articles concerning this topic [Deng et al., 2013], uses known variance reduction techniques, stratification and control variates, and combines this with a pre-experimental data to achieve variance reduction up to 50% for some metrics. The greatest variance reductions are achieved for metrics that varies significantly across the population and where the pre-experiment data is highly correlated with the metric of interest.

Experimental design, specifically repeated measures design, have also been discussed in the context of variance reduction [Guo and Deng, 2015]. The authors propose a framework that discusses different designs along with theoretical variance bounds and practical results from known distributions as well as real experiments. Some noteworthy properties of the proposed framework are that it does not impose independence restrictions on e.g. noise and do not assume that potential missing data have to be random. Results from real experiments have shown that up to 2/3 of the sample size can be reduced for some metric when applying this framework along with the previous mentioned variance reduction technique using pre-experiment data.

Other articles include an introduction to the notion of Overall Acceptance Criterion (OAC) that incorporates the OEC as well as statistical significance test to measure sensitivity [Drutsa et al., 2015]. Furthermore, improving sensitivity by accounting for delays in treatment effects have also been explored where it is shown that these modifications can improve the sensitivity for some loyalty metrics [Drutsa et al., 2017].

3.7 Randomization level

Randomization level or randomization unit refers to on what level the randomization algorithm assigns subjects to control and treatment, often on user or page view – level. The choice of randomization level is partly driven by what type of metrics are of interest but also the technical feasibility when it comes to the implementation of the randomization algorithm. Naturally, user-level metrics such as revenue per user, won't be measurable under page-view randomization [Deng et al., 2011]. Another important aspect to consider is the user experience where by design, the same user might be exposed to both control and treatment when randomization is done on page view level.

The statistical implications of the choice of randomization level concerns mainly variance estimation. It has been shown that page level randomization yields smaller variance estimation than user level randomization when analyzing page level metrics [Deng et al., 2011]. The randomization level will also affect the i.i.d assumption often assumed resulting in either underestimation or overestimation of the variance when using the standard empirical variance formula. Theoretical and practical results have been provided where a unified formula has been proposed where the authors also claims works when the randomization level is unknown [Deng et al., 2017].

3.8 Heterogenous Treatment Effects

Most causal inference frameworks are based on Average Treatment Effects (ATE), including e.g. frameworks that allows for optional stopping time as previously discussed. While ATE can answer questions such as "what metrics" are being effected and "how much", inferences about "who" they effect and "why" are harder to make, which warrants the use of models that accounts for Heterogenous Treatment Effects (HTE). For instance, how does treatment vary across demographics such as country, age and sex? Does a technical change effect different browsers or different mobile operating system differently? Perhaps certain user segments, such as top 1% spenders on an e-commerce site will react differently from the average spenders for a given treatment?

Articles have discussed different approaches to detect and account for HTE e.g. using nonparametric Bayesian Analysis [Taddy et al., 2016] and Total Variation Regularization [Deng et al., 2016b]. As previously mentioned, models controlling for FDR and accounting for the multiple comparison problem have also been proposed [Xie et al., 2018]. The last resource also includes a comprehensive discussion about previous work concerning HTE.

3.9 Metric design

The importance of good metrics is essential for successful experimentation. In a recent industry survey [Fabijian et al., 2018], among 44 respondents, only two respondents stated that their organization had successfully designed metrics that fully captured and were aligned with their business goals. The authors behind this survey also states that designing good metrics remains the biggest challenge for organizations who want to perform successful online experiments.

Two important qualities for metrics are the directionality and sensitivity of a metric [Deng and Shi, 2016]. The directionality refers to the degree of ambiguity of a metric. Take for instance the metric distinct queries/user. Greater value of this metrics implies higher user engagement which in many times is a desirable outcome. However, consider the metric in an online search setting, where a higher value might imply poor relevance of the search results. It is thus important to define metrics to accurately capture what one wishes to measure and complement metrics to each other to minimize ambiguity. Sensitivity of a metric is important since it implies faster detection of changes and thus cost savings when both required time and sample size are reduced when increasing sensitivity.

A comprehensive treatment of metric design is available [Deng and Shi, 2016], also examples of metric development frameworks from industry can be found [Dimitriev and Wu, 2016].

4 Discussion about RQ1

Note that the topics discussed until now are by no means an exhaustive compilation of previous research[5]. Also note that many of the sources referenced in the various sections also discusses many of the other topics in other sections where one particular source might not be referenced despite being relevant. Although not fully exhaustive and complete, the idea behind this exposition is to illustrate some underlying reasons as to why ROI calculations are challenging.

As seen in section 3, why ROI calculations are challenging and why for instance a 5% increase in conversion rate don't result in a 5% revenue increase can be due to several reasons; the practitioner can have misinterpreted the metric in question due to a statistical artifacts[6] such as Simpson's Paradox or the model can have failed to take into account important factors that have considerable effect on the OEC. It can of course also simply be the result of a random error inherent in all statistical tests. Although not discussed in length here, technical failures are often also a source of inaccurate test results.

Thus, it is rather straightforward as to why ROI calculations are challenging; we can't accurately compute what we can't accurately measure, and in the above sections we have illustrated some of the reasons to this inaccuracy and how to mitigate them.

5 Theory - The Risk Management Model

The following section provides a review of the statistical concepts utilized to build the model. The model itself will be described in Section 6.

5.1 Hypothesis Testing

Difference in means is usually tested with a two sample student's t-test and we will use this test to determine the difference between control and treatment. Since we are interested in knowing if control is better than treatment and to gauge the magnitude of the difference we define our null hypothesis and alternative hypothesis as:

$$H_0 : t \leq 0 \qquad H_1 : t > 0 \tag{7}$$

Let $\overline{X}_c$ and $\overline{X}_t$ denote the average treatment effect of control and treatment respectively, our t-statistics is defined as

$$t = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t}}} \tag{8}$$

[5]Some resources for reference: Comprehensive seminal work on A/B-testing Kohavi et al. [2009], A/B testing Online Encyclopedia [Kohavi and Longbotham, 2017] and a recent Mapping Study on the A/B research field [Ros and Runesson, 2018]

[6]An inference that does not reflect the real world due to bias in data collection, unintended consequences of measurement error or research design[OxfordIndex]

where n_i, σ_i denotes the sample size and standard deviation for control and treatment respectively.

5.2 Type I and Type II error

In statistical hypothesis testing, two types of inference errors are often of interest. Type I error refers to the error of incorrectly rejecting the null hypothesis even though it is true. Conversely, type II errors refers to the error of falsely accepting the null hypothesis given the alternative hypothesis being true. By convention, the probability of these errors are denoted α and β respectively. α is commonly known as statistical significance and $1 - \beta$ as statistical power.

	Type I and II error table	
	H_0 **True**	H_0 **False**
Accept H_0	Correct Inference True Negative Probability: 1-α	Type II error False Negative Probability: β
Reject H_0	Type I error False Positive Probability: α (stat. significance)	Correct Inference True Positive Probability: 1-β (stat. power)

Table 3: Visualization of the different errors along with the probabilities and associated terms

5.3 Sample Size Calculation

In the following section a sketch of the derivation of the sample size formula is provided. In order for us to commit a type I error when doing two sample, two side, student's t-test the following inequality must hold

$$t = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t}}} \geq t_\alpha \quad \implies \quad \overline{X}_t - \overline{X}_c \geq t_\alpha \sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t}} \tag{9}$$

Likewise, when a true difference Δ (absolute minimum detectable effect) is present, the following inequality holds when committing a type II error

$$\frac{\overline{X}_t - \overline{X}_c - \Delta}{\sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t}}} \leq -t_\beta \quad \implies \quad \overline{X}_t - \overline{X}_c \leq \Delta - t_\beta \sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t}} \tag{10}$$

From the two inequalities (9) and (10), we get

$$\Delta = (t_\alpha + t_\beta) \sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_t^2}{n_t}} \tag{11}$$

Let us now assume that $n_c = n_t = n_{ct}$ i.e. sample size in control and treatment are the same. Furthermore, since each observation in both control and treatment is treated as i.i.d. Bernoulli trials[7], in inequality (9) when H_0 is true, we have that $\sigma_c^2 = \sigma_t^2 = p_c(1-p_c)$ where p_c denotes the current conversion rate of control[8]. In inequality (10) under which H_1 is true, we have that $\sigma_c^2 = p_c(1-p_c)$ and $\sigma_t^2 = p_t(1-p_t)$. Thus, substituting and solving for n^* in (11) we get

$$\begin{aligned} \Delta &= t_\alpha \sqrt{\frac{2p_c(1-p_c)}{n_{ct}}} + t_\beta \sqrt{\frac{p_c(1-p_c)+p_t(1-p_t)}{n_{ct}}} \\ \implies n_{ct} &= (t_\alpha \sqrt{2p_c(1-p_c)} + t_\beta \sqrt{p_c(1-p_c)+p_t(1-p_t)})^2 \Delta^{-2} \end{aligned} \tag{12}$$

which is the formula used for optimal sample size calculation.

6 Risk Management Model

The model is inspired by a blog post[Georgiev, 2018] describing a commercially available software "ROI Calculator". Since no description of the math behind the software could be found, in conjunction with Conversionista, it was decided to purse this idea by creating a mathematical model based on the same idea and analyze the model.

The aim of the model is to provide a statistical tool to aid the A/B - testing practitioner to optimize sample sizes along with Type I & II error tolerance and minimum detectable difference. The main metric of interest is a version of a Return of Investment - ratio where the numerator is defined and quantified as expected uplift and the denominator as the expected loss based on a normal prior distribution specified by the practitioner. As such, the main metric of interest is conceptually defined as:

$$ExpROI = \frac{RiskAdjustedUplift}{RiskAdjustedCost} \tag{13}$$

The ratio can be thought of intuitively as a factor of how much larger our gains would have been compared to our costs if the same test with the same assumptions and parameters would be repeated many times. This is the concept which the model is inspired from, although in the blog post the quotient in question is referred to as a risk-reward ratio and is the inverse of equation 13. While an assumption about the normal distribution and a max effective size as presented in the following subsections are inspired from the blog post, the rest of the constructed model could not be found in existing literature.

The remainder of Section 6 is organized as follows:

[7] As explained in section 2.2

[8] Notice that the sample size formula is derived for $H_0 : t = 0$ i.e. $p_c = p_t$ and thus $\sigma_c^2 = \sigma_t^2$ for Bernoulli trials. We do this to avoid estimating p_t under our defined H_0 while we under H_1 estimate $p_t = p_c(1+\mu)$ where μ is defined in Section 6.2. Compared to our defined null hypothesis $H_0 : t \leq 0$, i.e. $\sigma_c^2 \geq \sigma_t^2$ when $p_c < 0.5$, this will yield slightly more conservative sample size and slightly underestimate the sample size when $p_c > 0.5$ i.e. $\sigma_c^2 \leq \sigma_t^2$.

- 6.1 Model Assumptions describe what assumption goes into the model
- 6.2 Prior Normal Distribution explains the idea of which the practitioner incorporate his subjective beliefs of the test outcome into the model
- 6.3 Effective Size describes how long the effect of any positive test will last
- 6.4 ExpROI specification describes the mathematical components of the quotient
- 6.5 Fixed Cost and Revenue per User describes a simple extension the description in 6.4

6.1 Model Assumptions

As we have seen, there are several different monitoring and measuring schemes in use among practitioners. For this reason, we chose to build our model on the assumption that a classical fixed horizon testing scheme is used, i.e. that sample size along with statistical significance, power and MDE is chosen in advanced and not altered or adjusted during the course of the test. The fixed horizon testing scheme is arguably the scheme most widely in use as well as the least controversial, thus providing a good starting point for the proposed model. Furthermore we make the following assumptions in order to start with a model that is not too complex:

- We test for average treatment effects, that is any heterogeneous treatment effects (see 3.8) are not accounted for
- Visitors are treated as i.i.d. random variables - that is we assume no interaction between the users
- Revenue per user (RPU) is constant and only conversion rate changes when there is a difference between control and treatment
- We assume one control and one treatment group

6.2 Prior Normal Distribution

The first step of the model is for the practitioner to provide a subjective prior normal distribution on relative uplift of the conversation rate, that is the quantity

$$U = \frac{p_t - p_c}{p_c} \sim N(\mu, \sigma) \tag{14}$$

where p_i is the conversion rate for treatment, $i = t$, and control, $i = c$.

Furthermore, note that the minimum detectable effect in the sample size formula in equation (12) is the absolute difference, define δ as the relative minimum detectable effect and we have

$$\Delta = p_t - p_c = \delta p_c \tag{15}$$

The observant reader will notice that with the notation we have used we have seemingly defined $U = \delta$. However, there is an important difference. While U denotes the random variable for relative uplift that we assign a subjective normal distribution to, δ denotes the relative minimum detectable effect which is a fixed testing parameter value. Admittedly, this is an unfortunate use of notation however there should be no confusion after this clarification.

The distribution in equation 14 reflects the practitioner's beliefs of the particular experiment to be conducted and can either be based on expertise, historical test results of similar experiments or pilot tests. For instance, one industry vendor have published a paper [Qubit, 2017][9] where data from 6700 large scale online experiments are aggregated to estimate normal distribution parameters for a number of categories such as "up-sell", "page redesign", "free delivery", "landing page" etc.

6.3 Effective size

When a successful test is conducted and changes implemented, the observed difference between control and treatment is not always sustained over longer period of times. For instance, a 5% site wide conversion rate increase rarely results in a permanent 5% sales increase over time. To account for this, we model this by letting the practitioner set a max size n_{max}. Furthermore, define n_e as the number of visitors that will be affected by implemented changes post testing.

$$n_e = n_{max} - 2n_{ct}$$

Clearly, the sample size for control and treatment will always be $2n_{ct} \leq n_{max}$ thus setting a constraint for the testing sample sizes when optimizing ExpROI, see Figure 4.

n_{max} can be interpreted in a number of different ways. For instance, setting n_{max} to one year traffic to the website can be interpreted as any potential changes made will only have a constant effect over a year starting from the testing period or alternatively, having a two year effect with any effect decreasing linearly to zero over time. Furthermore, it also allows the practitioner to incorporate any future growth (or decrease) in number of visitors by varying n_{max}.

[9] Independently assured by PwC

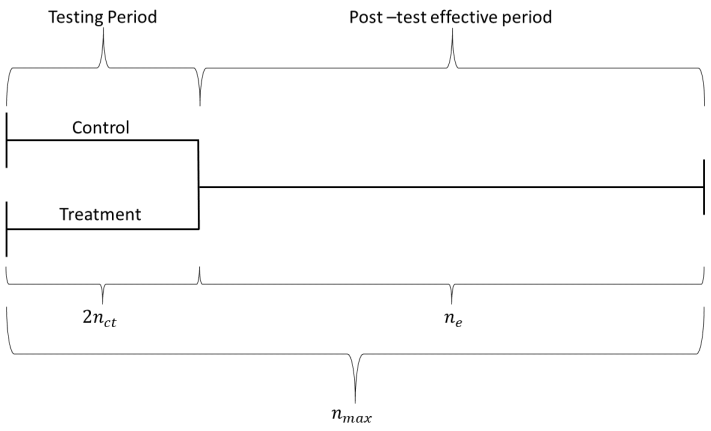

Figure 4: Visualization of the testing period and post - test effective test and the corresponding sample sizes to each period.

6.4 ExpROI specification

In this section we define the numerator and denominator in ExpROI. In the case of testing, the expected uplift will consist of three terms while the expected cost will consist of two terms.

$$ExpROI = \frac{RiskAdjustedUplift}{RiskAdjustedCost} = \frac{R_1 + R_2 + R_3}{C_1 + C_2} \tag{16}$$

R_1, R_2 and R_3 denotes the contribution from a True Positive result, a False Negative result and an undetectable difference respectively. Conversely, C_1 and C_2 denote the contribution from a False Positive result and a True Negative result respectively. We use the following notation in this section:

Let $\Phi(x)$ denote the CDF for the standard normal and $\phi_{\mu,\sigma}(x)$ the PDF for a normal distribution with parameters μ and σ.

$$P(H_1) = 1 - \Phi(\frac{-\mu}{\sigma})$$

$$P(H_0) = 1 - P(H_1)$$

$$P(t > \delta) = 1 - \Phi(\frac{\delta - \mu}{\sigma})$$

$$E_{a<t<b}(U|H_1) = \frac{\int_a^b x\phi_{\mu,\sigma}(x)dx}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

Recall that U denotes the random variable that the practitioner assigns a subjective normal to, H_0 and H_1 the null hypothesis and the alternative hypothesis respectively.

6.4.1 Risk adjusted Uplift

The formula for risk adjusted uplift when implementing immediately without testing is fairly straightforward.

$$RiskAdjustedUplift = P(H_1)E(U|H_1)n_{max} \tag{17}$$

Some corrections to equation 17 is made in the case of testing to account for the statistical power and minimum detectable effect:

1. **True Positive** - here we account for the probability of the true uplift being greater than the minimum detectable difference as well as accounting for the power. This effect applies to $n_e + n_{tc}$ since it will be implemented if and when detected. Since we will never reap the benefits from a positive change below δ after implementation, our expectation will have δ as lower bound.

$$\begin{aligned} R_1 &= P(t > \delta, H_1)(1-\beta)E_{\delta<t}(U|H_1)(n_e + n_{tc}) && (18)\\ &= P(H_1)P(t > \delta|H_1)(1-\beta)E_{\delta<t}(U|H_1)(n_e + n_{tc}) && (19)\\ &= P(H_1)\frac{P(t > \delta)}{P(H_1)}(1-\beta)E_{\delta<t}(U|H_1)(n_e + n_{tc}) && (20)\\ &= P(t > \delta)(1-\beta)E_{\delta<t}(U|H_1)(n_e + n_{tc}) && (21) \end{aligned}$$

   In other words, this is the contribution from a test outcome where treatment truly is better than control and we are able to detect it.

2. **False Negative** - when a true uplift greater than δ is not detected, it implies that the test statistics given, i.e. the observed difference in cumulative conversion rate has not been big enough relative to δ. However, benefits during testing period has most probably occurred. We model this by taking the expectation over the positive support up to δ as well as accounting for type II error probability and applying the effect to the test sample size since it will not be implemented.

$$R_2 = P(t > \delta)\beta E_{0<t<\delta}(U|H_1)n_{tc} \tag{22}$$

3. **Undetectable difference** - in case of the true uplift lying between zero and δ, the effect will not be detected and thus control will be kept. However, being a positive uplift still, the effect during testing is accounted for in down below formula.

$$R_3 = P(0 < t < \delta)E_{0<t<\delta}(U|H_1)n_{tc} \tag{23}$$

With the three terms above, we have specified the numerator in our key metric, accounting for all events where a the net revenue is positive compared to status quo as well as adjusting for the probability of the events.

6.4.2 Risk adjusted Costs

Analogously to the risk adjusted uplift, the risk adjusted costs for implementing immediately without testing becomes:

$$RiskAdjustedCost = P(H_0)E(U|H_0)n_{max} \tag{24}$$

The risk adjusted cost associated with testing is also corrected, this time for statistical significance. Note that assigning a prior normal centered in zero, that is that uplift and loss is equally likely, would result in an ExpROI of one when implementing immediately without testing.

1. **False Positive** - this term quantifies the risk of implementing a treatment worse than control. The statistical significance is accounted for and the effected size is all visitors.

$$C_1 = P(H_0)\alpha|E(U|H_0)|(n_e + n_{tc}) \tag{25}$$

2. **True negative** - the cost associated with correctly identifying a bad treatment comes from the testing period when testing the bad treatment.

$$C_2 = P(H_0)(1-\alpha)|E(U|H_0)|n_{tc} \tag{26}$$

Thus, the risk adjusted costs for testing become the sum of the above two terms. The absolute sign of the expectation in equation (25) and (26) respectively is simply to get the cost and thereby the ratio positive.

6.5 Fixed Costs and Revenuer per User

Conducting tests will of course incur other costs than the two risk adjusted cost term described above. Such costs can for instance be costs related to setting up the infrastructure needed for testing or the payroll for the developers and data scientists working with the tests. How to accurately estimate these kind of fixed costs is outside the scope of this thesis but given a fixed cost we can incorporate the costs into the formula. Note that revenue per user (RPU), defined as the mean of revenue per user, have to be included in the formula where it otherwise would be canceled out when no fixed costs is assumed.

$$ExpROI = \frac{(R_1 + R_2 + R_3)RPU}{(C_1 + C_2)RPU + C_f} \tag{27}$$

7 Analysis

In this section we perform exploratory analysis of the model, starting with the case of varying one variable and fixing the others and progressively varying more variables. In each addition of a new varying variable, the function is optimized again without any results from past optimizations carrying over. Lastly, we add the fixed cost parameter and examine the effect of fixed costs compared to no fixed cost term.

7.1 Optimal MDE

In this section we analyze the optimum of the minimum detectable difference (MDE), denoted δ throughout this project, under what is arguably standard default values for statistical significance and power, $\alpha = 0.05, \beta = 0.2$.

In Figure (5), we notice how some graphs are undefined on some parts of the domain. This is due to the fact that for the given parameters values, at those values of δ, the sum of the sample size of control and treatment exceeds the maximum effective size n_{max}, rendering a test undefined for the given effective time in the model. Thus, many of the optimal MDE is quite large relative to the standard deviation, many times the MDE is a factor of one to two of the standard deviation. This implies that or ability to detect a true difference suffers since the true difference would have to be larger than MDE, with a large MDE the probability of the true difference being greater than the MDE is low.

Furthermore, we see how maximum ExpROI often is reached around roughly 1-5% MDE when a maximum is within the visible domain suggesting (wrongfully) a fairly robust range of values for δ. Comparing subplot 4 and 5 we can see how a nonchalant selection of the value of δ can completely undermine a test's legitimacy. With assumed parameters in subplot 5, the optimum MDE is just shy of 5%. A 5% in the setting of subplot 4 however, would give us a value of less than 0.5, certainly less than 1, suggesting that our expected losses would be greater than any expected uplift. On the other hand, the optimal δ in subplot 4 around 2%, runs the risk of resulting in a sum of testing sample sizes bigger than n_{max} in the setting of subplot 5.

Figure 5: Plotting ExpROI against $0 < \delta < 0.2$ for a number of values for μ, σ, n_{max} and p_c. Depending on service and industry, different control conversions p_c can be seen as typical conversion rates. Here we different control conversion rates, 0.05 & 0.7, to illustrate the difference.

Other noteworthy insights are how optimal MDE relates to the variance of the relative increase in uplift. We can consistently see, for instance comparing subplot 1 to 5 or 13 to 17, that under a greater assumed variance ceteris paribus, the optimal MDE becomes larger. This matches our intuition; if there is any deviations from zero, those differences are bound to be larger on average the greater the variance is, allowing for less strict and thus larger MDE.

Lastly, focusing on subplot 13 - 20, that is when $\mu < 0$, the range of most of these plots are below one suggesting that conducting tests might not be worth it, which makes sense in a business setting - why test something you believe have greater probability to fail than succeed? However, if either the variance σ^2, maximum effective size n_{max} or baseline p_c or a combination of these are large enough, our ExpROI may exceed one as seen in Figure (6) , suggesting that doing tests might be beneficial under these assumptions.

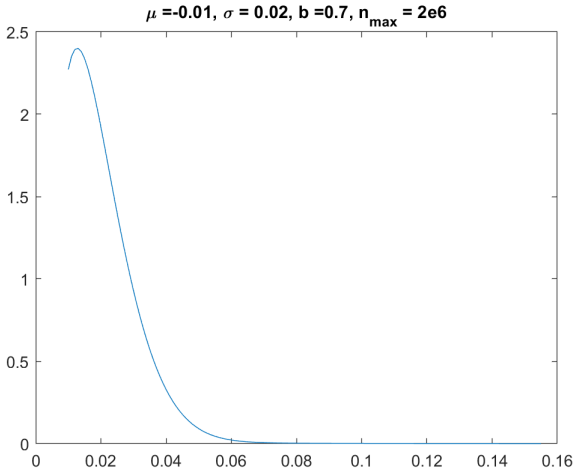

Figure 6: Negative μ with increased σ and n_{max}

7.2 Optimal α and β

To get a sense of how ExpROI behaves while varying α and β, we fix δ. These $\delta : s$ are chosen based on the optimal MDE in Figure (5). Letting the x- and y-axis represent α and β respectively, and the z-axis represent ExpROI we get the following surfaces for some chosen parameter values.

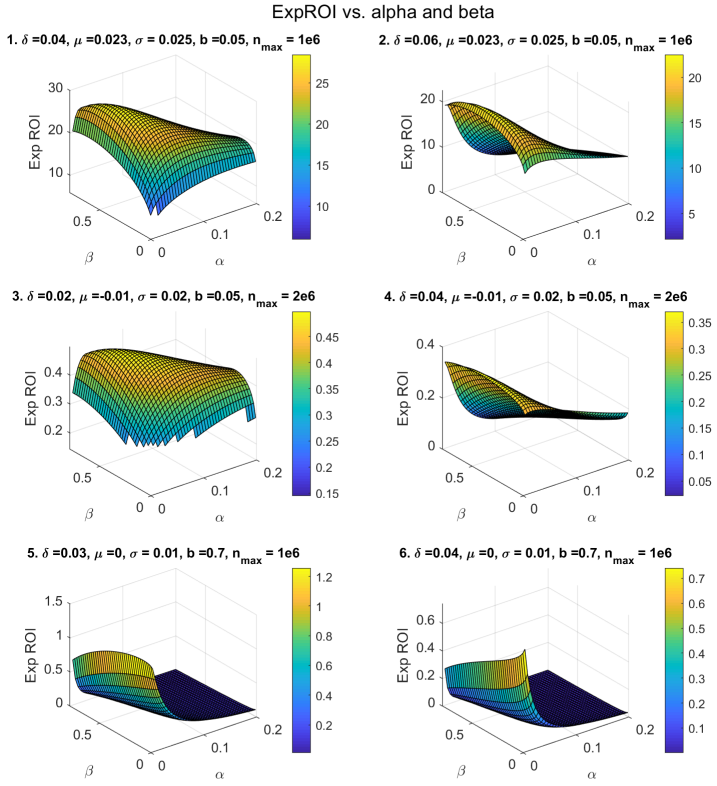

Figure 7: ExpROI vs. α and β for some selected parameter values

In our plots, we see that in our visible domain a global maximum exist. Furthermore, ExpROI seems to premier large β for small control conversion rate, far larger than what is usually accepted in statistical tests as seen in Table 4. Also, for large control conversion rates, drastically smaller sample sizes are needed and high statistical significance can be achieved relatively cheaply as indicated by the low α.

Subplot	Parameters					Results			
	δ	μ	σ	p_c	n_{max}	ExpROI	α_{opt}	β_{opt}	n_{ct}
1.	0.04	0.023	0.025	0.05	1e6	28.4506	0.0400	0.4700	79250
2.	0.06	0.023	0.025	0.05	1e6	22.4815	0.0150	0.4100	60840
3.	0.02	-0.01	0.02	0.05	2e6	0.4977	0.0600	0.5300	207850
4.	0.04	-0.01	0.02	0.05	2e6	0.3700	0.0100	0.3700	168210
5.	0.03	0	0.01	0.7	1e6	1.2557	0.0050	0.1100	13680
6.	0.04	0	0.01	0.7	1e6	0.7411	0.0050	0.0100	12700

Table 4: Results of optimized ExpROI for α and β with fixed δ. Note that the surface function is evaluated at discrete increments of α and β, 0.005 and 0.02 respectively.

The parameter values are chosen so that the only difference between subplot 1 and 2, 3 and 4, 5 and 6 respectively, is the MDE. We notice how both α and β decreases as MDE increases. This is explained by as our MDE increases, our ability to detect true difference suffers. To compensate for this, α decreases i.e. the statistical significance increases so that the false positive term (25) is given less weight and the true negative term (26) more weight. Likewise, a decrease in β results in increased power and thus less weight in false negatives (22) and increased weight in true positive (18). To summarize; tests with less sensitivity are compensated with more precision.

7.3 Optimizing for all covariates

We investigate the existence of a global maximum visually by plotting a 3d - heatmap letting the three axis represent α, β and δ respectively and introduce the fourth dimension, ExpROI, as the colour of the plot.

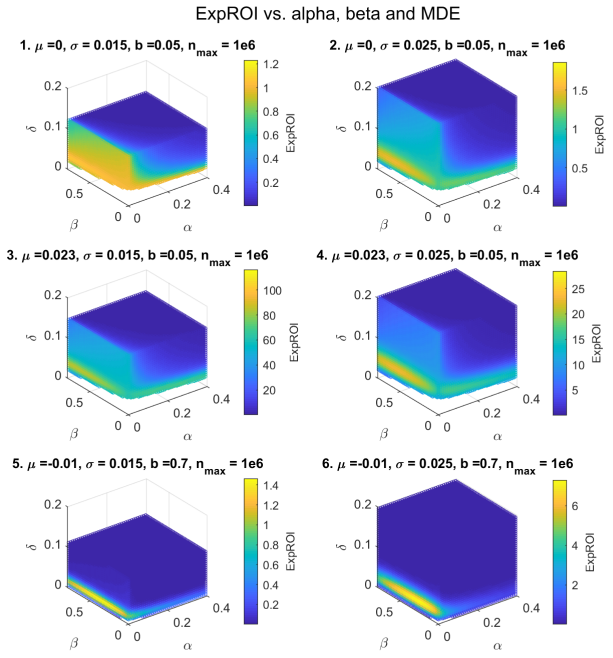

Figure 8: ExpROI vs. α, β and δ for some selected parameter values

As seen in Figure (8), from the the looks of it, it seems that a global maximum exists when optimizing over our covariates for the given parameter values. To conclusively prove the existence of a global maximum, one would have to prove it analytically. While such a derivation will not be explored in this project, based on the plots and our intuition, we claim it is quite likely that global maximum exists for all parameter values.

Subplot	Parameters				Results				
	μ	σ	p_c	n_{max}	ExpROI	α_{opt}	β_{opt}	δ_{opt}	n_{ct}
1.	0	0.015	0.05	1e6	1.2305	0.0850	0.5500	0.0200	147480
2.	0	0.025	0.05	1e6	1.8619	0.0550	0.5300	0.0300	97860
3.	0.023	0.015	0.05	1e6	116.3298	0.0550	0.5500	0.0300	91440
4.	0.023	0.025	0.05	1e6	28.4414	0.0350	0.5100	0.0400	75810
5.	-0.01	0.015	0.7	1e6	1.4652	0.0050	0.3900	0.0150	31020
6.	-0.01	0.025	0.7	1e6	7.2916	0.0050	0.2700	0.0250	13900

Table 5: Results of optimized ExpROI for α, β and δ.

Barring the insights we gained from the previous two sections, we see how small changes in our parameter values have drastic effects on our results. Comparing subplot 1 and 2 in Table 5, ExpROI indicates that a test is barely worth

it whereas performing tests under the same assumptions with greater variance as seen in the results of subplot 2 yields more promising indications.

We can also see how a greater variance implies less benefits of test when $\mu > 0$ as seen in subplot 3 and 4 in Table 5, which is expected since $P(t < 0)$ increases as σ increases ceteris paribus and vice versa. Naturally, it makes sense intuitively too since a positive expected uplift paired with a low uncertainty means less need for testing.

7.3.1 Sensitivity Analysis

As seen in Figure (9), which is a zoomed in and rotated version of Figure (8), a large yellow band can be identified in the plots indicating similar values of ExpROI over the volume of the yellow sections. Using a built-in numerical optimization routine in matlab[10], we maximize ExpROI for four different cases for each of the subplots to examine the sensitivity of ExpROI. The results are presented in Table 6.

As we can see from our results, for some parameter values, standard values for the statistical significance and power yields ExpROI values close to the numerically optimized value. Furthermore, fixing β to 0.2 and optimize for α and δ yields values of ExpROI that is greater than 90% of the global maximum in all of our tested cases. This raises the question whether ExpROI is a reliable metric and what it really means to maximize this quotient.

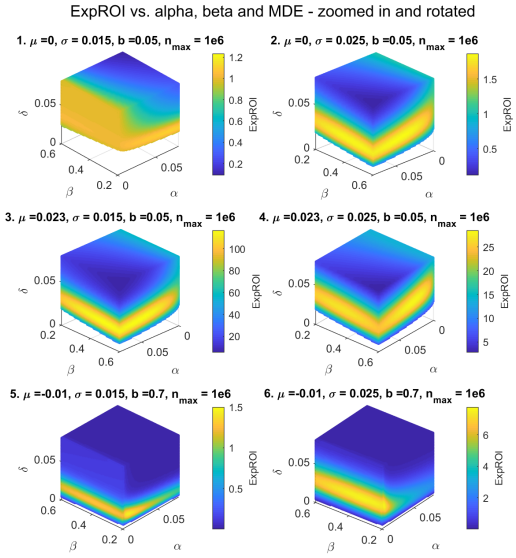

Figure 9: Zoomed in and rotated version of Figure 8

[10]The routine in question is fminunc() which uses the BFGS Quasi-Newton to optimize the function [MathWorks]

Subplot	Fixed		Results				
$(\mu, \sigma, p_c, n_{max})$	α	β	ExpROI	α	β	δ	ExpROI % of Max
	-	-	1.237	0.102	0.574	0.018	100%
1.	-	0.2	1.123	0.185	-	0.020	90.8%
(0, 0.015, 0.05, 1e6)	0.05	-	1.212	-	0.646	0.021	98.0%
	0.05	0.2	1.09	-	-	0.028	88.1%
	-	-	1.864	0.052	0.518	0.031	100%
2.	-	0.2	1.677	0.08	-	0.034	93.0%
(0, 0.025, 0.05, 1e6)	0.05	-	1.864	-	0.522	0.031	99.4%
	0.05	0.2	1.656	-	-	0.037	92.8%
	-	-	116.76	0.062	0.544	0.028	100%
3.	-	0.2	102.28	0.115	-	0.030	90.3%
(0.023, 0.015, 0.05, 1e6)	0.05	-	116.28	-	0.576	0.029	99.9%
	0.05	0.2	97.193	-	-	0.035	88.2%
	-	-	28.486	0.036	0.486	0.041	100%
4.	-	0.2	25.729	0.054	-	0.043	90.3%
(0.023, 0.025, 0.05, 1e6)	0.05	-	28.217	-	0.446	0.039	99.1%
	0.05	0.2	25.717	-	-	0.044	90.3%
	-	-	1.500	0.007	0.350	0.016	100%
5.	-	0.2	1.447	0.008	-	0.016	96.5%
(-0.01, 0.015, 0.7, 1e6)	0.05	-	1.044	-	0.203	0.012	69.6%
	0.05	0.2	1.044	-	-	0.012	69.6%
	-	-	7.756	0.002	0.302	0.029	100%
6.	-	0.2	7.612	0.002	-	0.030	98.1%
(-0.01, 0.025, 0.7, 1e6)	0.05	-	3.231	-	0.114	0.017	41.7%
	0.05	0.2	3.162	-	-	0.016	40.1%

Table 6: Numerical optimization of six set of parameter values with four cases in each set.

7.4 Simulations for β

As we have seen in above sections, around the maximum value of ExpROI, the changes along β is relatively small as seen in the relative flat shape along the β - axis in Figure (7) and the elongated yellow part in Figure (9). This warrants us taking a closer look on β. For the same parameters combination as Table 6 we first find the optimal α and δ along with β. We then vary β between 0.1 - 0.6 with 0.1 increments and simulate 10,000 iteration for each combination. Lastly, we sort the result of each iteration for each combination and plot the results.

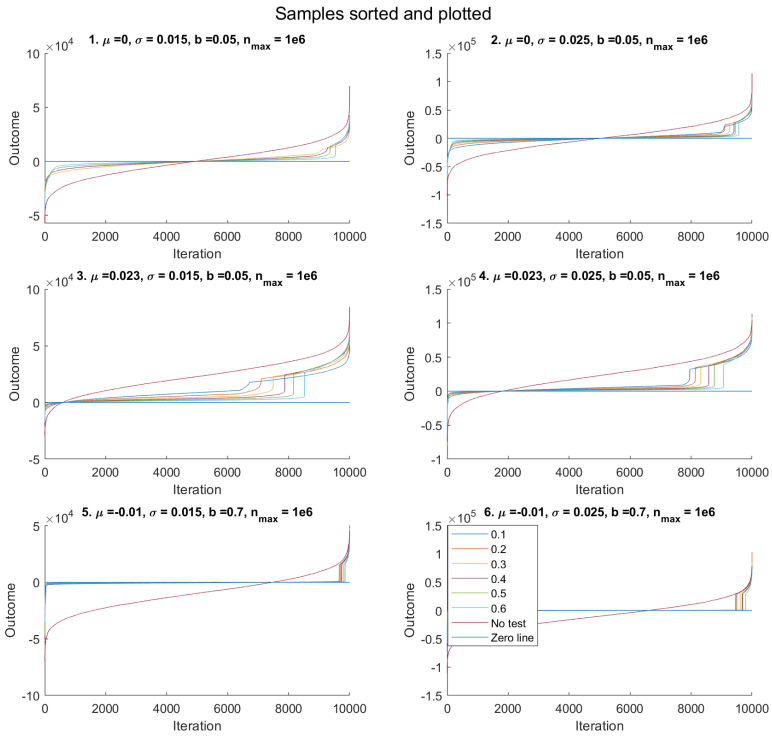

Figure 10: Iterations on x-axis and outcome on y-axis. Legend can be found in bottom right plot.

As seen from Figure 10 and Table 7, the number of times we win are fairly consistent over the different β i.e. conducting test with different values for the statistical power will not increase our likelihood to win. It does however as expected effect how often we correctly choose a better treatment. The jumps in the graph corresponds to the contribution of true positive R_1, equation (18), while the modest upwards slopes are contributions from R_2 and R_3, equation (22) and respectively (23). Empirical density functions corresponding to the six cases can be found in appendix A.

7.5 Fixed Cost ExpROI

Here we add a fixed cost parameter as well as including revenue per user as defined in equation (27). Adding positive RPU and fixed cost term to the quotient will add more weight to the denominator compared to the vanilla case. To offset this, the relative value of the numerator has to increase as well. The most straightforward way to do this is to increase the weight of R_1, i.e. decrease

β. Looking at equation (27), we hypothesize that the more dominating C_f is relative the other terms in the quotient, the greater the beta reduction is.

To examine this we evaluate and optimize 100 combinations of fixed cost and RPU where fixed cost ranges from 10,000 to 100,000 and RPU 100 to 1,000 for the same set of parameter values as in Table 6. A visualization of the changes is included in Figure 11.

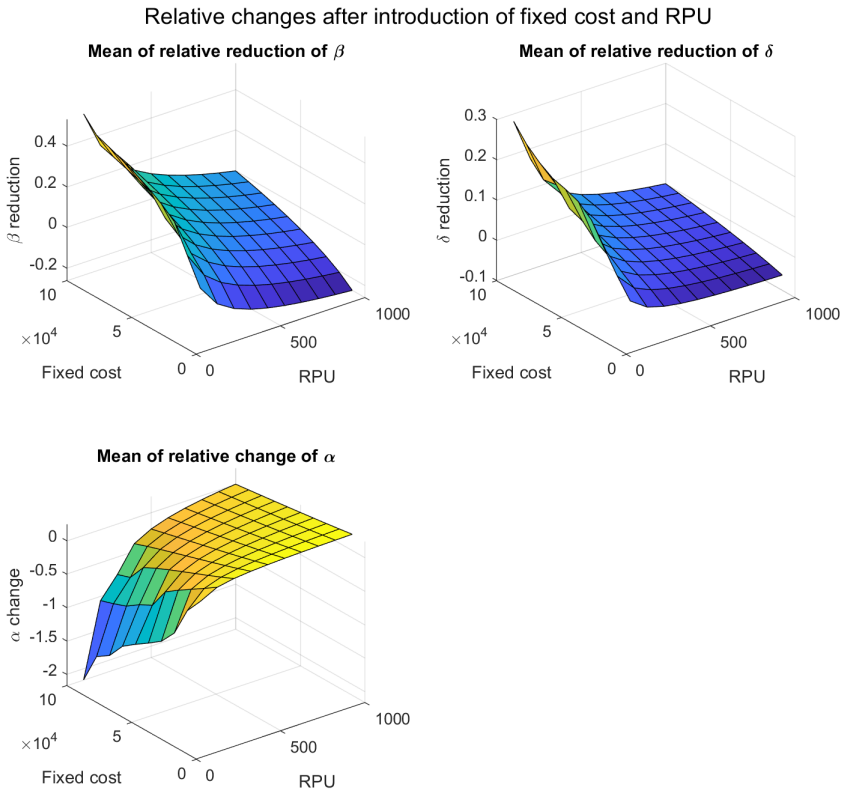

Figure 11: Relative changes when optimizing ExpROI after introduction of RPU and fixed cost for the same set of parameter values as Table 6. Note that the graphs are only indicative since the graph shows the mean changes of only six parameter combinations.

.

Not only do β decrease but we see a reduction in δ as well as relatively large increase in α. With other words, we managed to increase the statistical power and the sensitivity by introducing a large fixed cost and trading statistical significance; there is no such thing as a free lunch. See appendix B for table of

full results for some selected parameter values.

Subplot $(\mu, \sigma, p_c, n_{max})$	Results	β 0.1	0.2	0.3	0.4	0.5	0.6
1. (0, 0.015, 0.05, 1e6)	ExpROI	NA	NA	1.1497	1.2715	1.3970	1.3504
	% of Max	NA	NA	82.3%	91.0%	100%	96.7%
	Win Rate	NA	NA	50.5%	50.1%	49.9%	50.2%
2. (0, 0.025, 0.05, 1e6)	ExpROI	1.2881	1.6804	1.8129	1.8471	2.0800	2.1320
	% of Max	60.4%	78.8%	85.0%	86.6%	97.6%	100%
	Win Rate	49.2%	50.3%	50.5%	50.3%	50.7%	50.1%
3. (0.023, 0.015, 0.05, 1e6)	ExpROI	76.0867	92.5744	112.6850	125.9870	147.1618	128.5313
	% of Max	51.7%	62.9%	76.6%	85.6%	100%	87.3%
	Win Rate	93.8%	93.6%	93.8%	93.9%	94.2%	93.8%
4. (0.023, 0.025, 0.05, 1e6)	ExpROI	20.7896	24.8373	27.6183	32.8742	31.5968	25.4261
	% of Max	63.2%	75.6%	84.0%	100%	96.1%	77.3%
	Win Rate	82.3%	81.4%	81.5%	82.2%	83.0%	82.0%
5. (-0.01, 0.015, 0.7, 1e6)	ExpROI	1.2313	1.4205	1.4797	1.6193	1.3709	1.2788
	% of Max	76.0%	87.7%	91.4%	100%	84.7%	79.0%
	Win Rate	24.1%	24.9%	25.2%	25.4%	25.0%	25.7%
6. (-0.01, 0.025, 0.7, 1e6)	ExpROI	7.4330	8.4925	7.5939	7.1491	8.3816	5.6299
	% of Max	87.5%	100%	89.4%	84.2%	98.7%	66.3%
	Win Rate	35.3%	35.2%	34.4%	34.3%	34.3%	34.1%

Table 7: Results from simulations. 10,000 iterations run for each parameter combination. Note that the maximum value from the simulations are greater than those found in table 6. This is because we correct for a lower mean in our formula given a false negative (equation 22) whereas the simulation simulates false negatives invariant to the size of the uplift, as long as it is positive, thus resulting in slightly higher values of ExpROI.

8 Model Discussion

The aim with the model was to investigate whether the model could provide practitioners an easy way to gauge the financial viability of conducing test based on certain beliefs about the outcome and if test parameters could be optimized accordingly. The entire model rests on the assumption that the difference between the conversion rate in control and treatment can be described by a normal distribution. Firstly, there is the question whether such a assumption is valid at all. Secondly, suppose the difference actually can be accurately described by a normal distribution, the question remains how to accurately specify the parameters to the distribution. Despite this, an argument to why such an assumption is viable is to compare it to the alternative; if no other models and methods are readily available, an informed guess might be a better option than blindly choosing test parameter values or revert to default ones.

From our exploratory analysis we have identified and characterized certain behaviour of ExpROI. We have also shown how ExpROI are unable to account for one important factor. The magnitude of loss and wins is something ExpROI is naturally agnostic to it being a quotient; it does not discriminate between 1:10 and 10:100. With other words, it does not optimize the frequency of which we are able to correctly identify a better treatment. This effect results in high

β and δ when maximizing the quotient. We can increase power and sensitivity to some degree by introducing a fixed cost and trading α, still, this does not entirely address the aforementioned issue. A possible remedy to this would be to include complimentary metrics that measures mean or the rate of which a positive treatment is correctly identified. However, this would most likely decrease the ease of use and interpretability of the model when introducing additional metrics. To this end, to blindly maximize ExpROI as we have done in this project would not be advisable. However, as a standalone metric, it could be used as a sanity check to see if a with some chosen test parameters are expected to have higher possibility of success than failure by keeping the quotient above one.

9 Further Research

As discussed above, the immediate extension would be to include additional metrics to account for the magnitude and frequency of wins. If such a model is proven to viable there are many venues to explore as seen in the literature review.

A possible extension of the model is to include more treatments into the model as well as model the dependencies of tests that follow each other sequentially in a customer buying journey. For instance, when running concurrent tests on the landing page as well as the checkout page, the outcome of the the test on the landing page will undeniably affect the test on the checkout page. A second natural extension, is to model changes in revenue distribution as opposed to assuming that treatments only affect conversion rate.

Other possible areas to explore includes taking the general idea of ExpROI, along with its' complimentary metrics, and applying it on statistical frameworks that allows continuous monitoring or detection of heterogeneous treatment effects. This way, parameters in such models can be tuned based on financial sound metrics.

Lastly, it can be interesting to properly examine the fixed cost term in the formula or add other relevant terms both in the numerator and the denominator. One could easily argue for instance, that there is inherent value of gaining insights when conducting tests e.g. that practitioner can infer certain customer behaviours from tests that are successful. Other costs that might be relevant to include could be costs accounting for buggy implementation or other technical failures.

References

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for experimenters*. John Wiley & Sons, Hoboken, 2 edition, 2005.

Thomas Crook, Brian Frasca, Ron Kohavi, and Roger Longbotham. Seven pitfalls to avoid when running controlled experiments on the web. In *KDD*, Paris, France, 2009.

Alex Deng and Xiaolin Shi. Data-driven metric development for online controlled experiments: Seven lessons learned. In *KDD*, San Fransisco, USA, 2016.

Alex Deng, Roger Longbotham, Toby Walker, and Ya Xu. Choice of the randomization unit in online controlled experiment. 2011.

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM*, Rome, Italy, 2013.

Alex Deng, Tianxi Li, and Yu Guo. Statistical inference in two-stage online controlled experiments with treatment selection and validation. *Proceeding WWW '14 Proceedings of the 23rd international conference on World wide web*, pages 609–618, 2014.

Alex Deng, Jiannan Lu, and Shouyuan 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)*, Montreal, Canada, 2016a.

Alex Deng, Pengchuan Zhang, Shouyuan Chen, Dong Woo Kim, and Jiannan Lu. Concise summarization of heterogeneous treatment effect using total variation regularized regression. 2016b.

Alex Deng, Jiannan Lu, and Jonathan Litz. Trustworthy analysis of online a/b tests: Pitfalls, challenges and solutions. In *WDSM*, Cambridge, United Kingdom, 2017.

Pavel Dimitriev and Xian Wu. Measuring metrics. In *CIKM*, Indianapolis, USA, 2016.

Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. A dirty dozen: Twelve common metric interpretation pitfalls in online controlled experiments. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1427–1436, 2017.

Alexey Drutsa, Anna Ufliand, and Gleb Gusev. Practical aspects of sensitivity in online experimentation with user engagement metrics. In *CIKM*, Melbourne, Australia, 2015.

Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. Using the delay in a treatment effect to improve sensitivity and preserve directionality of engagement metrics in a/b experiments. In *International World Wide Web Conference Committee*, Perth, Australia, 2017.

Aleksander Fabijian, Pavel Dmitriev, Helena Holmström Olsson, and Jan Bosch. Online controlled experimentation at scale: An empirical survey on the current state of a/b testing. In *SEAA*, Prague, Czechia, 2018.

Georgi Georgiev. Risk vs. reward in a/b tests: A/b testing as risk management. 2018. URL `http://blog.analytics-toolkit.com/2017/risk-vs-reward-ab-tests-ab-testing-risk-management/`. last accessed 06-06-2019.

Yu Guo and Alex Deng. Flexible online repeated measures experiment. 2015.

Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the long-term: It's good for users and business. In *Proceeding KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1849–1858, Sydney, NSW, Australia, 2015.

Ramesh Johari, Pete Koomen, Lenoid Pekelis, and David Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *KDD Applied Data Science Paper*, Halifax, NS, Canada, 2017.

Nianqiaio Ju, Diane Hu, and Adam Henerson Liangjie Hong. A sequential test for selecting the better variant. In *Association for Computing Machinery*, Melbourne, Australia, 2019.

Ron Kohavi and Roger Longbotham. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter*, 12(2):31–35, 2010.

Ron Kohavi and Roger Longbotham. Online controlled experiments and a/b testing. In C Sammut, editor, *Encyclopedia of Machine Learning and Data Mining*. Springer, Boston, 2017.

Ron Kohavi and Stefan Thomke. The surprising power of online experiments. *Harvard Business review*, September - October, 2017. doi: 3245467.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data and Mining Discovery*, 2009.

Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *KDD*, Beijing, China, 2012.

Luo Lu and Chuang Liu. Separation strategies for three pitfalls in a/b testing. *UEO Workshop, KDD. ACM, 2014*, 2014.

MathWorks. fminunc. URL https://se.mathworks.com/help/optim/ug/fminunc.html. Reference entry, last accessed 06-06-2019.

OxfordIndex. artefacts, statistical and methodological. *A Dictionary of Sociology*. URL https://oxfordindex.oup.com/view/10.1093/oi/authority.20110803095426317. Reference entry, last accessed 06-06-2019.

Qubit. What works in e-commerce - a meta-analysis of 6700 online experiments. 2017. URL https://www.qubit.com/wp-content/uploads/2017/12/qubit-research-meta-analysis.pdf. last accessed 02-03-2019.

Rasmus Ros and Per Runesson. Continuous experimentation and a/b testing: A mapping study. In *2018 ACM/IEEE 4th International Workshop on Rapid Continuous Software Engineering*, 2018.

Matt Taddy, Matt Gardner Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. In *American Statistical Association Journal of Business Economic Statistics*, volume 344, 2016.

Robert E. Wheeler. Portable power. *Technometrics*, 16(2):193–201, 1974.

Andrew R. William and Andrew H. Briggs. *Statistical Analysis of Cost - effectiveness Data*. John Wiley & Sons, West Sussex, 2006.

Yuxiang Xie, Nanyu Chen, and Xiaolin Shi. False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments. In *KDD*, New York, USA, 2018.

Ya Xu, Nanyu Cheng, Adrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/b testing challenges in large scale social networks. In *KDD*, Sydney, Australia, 2015.

Appendices

A Empirical density functions

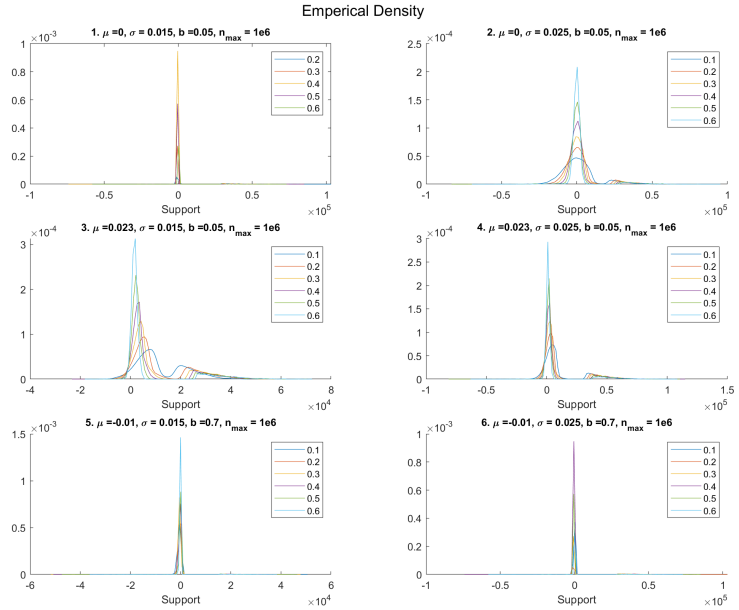

Figure 12: Empirical density function corresponding to the simulation run in Table 7

B Results after introducing fixed cost and RPU

Subplot	Fixed		Results				
$(\mu, \sigma, p_c, n_{max})$	α	β	ExpROI	α	β	δ	ExpROI % of Max
	-	-	1.155	0.131	0.504	0.017	100%
1.	-	0.2	1.074	0.211	-	0.019	93.0%
(0, 0.015, 0.05, 1e6)	0.05	-	1.115	-	0.568	0.022	96.5%
	0.05	0.2	1.036	-	-	0.027	89.7%
	-	-	1.746	0.060	0.470	0.030	100%
2.	-	0.2	1.606	0.086	-	0.033	92.0%
(0, 0.025, 0.05, 1e6)	0.05	-	1.742	-	0.487	0.031	99.8%
	0.05	0.2	1.580	-	-	0.036	90.4%
	-	-	54.969	0.206	0.245	0.023	100%
3.	-	0.2	54.752	0.226	-	0.023	99.6%
(0.023, 0.015, 0.05, 1e6)	0.05	-	49.883	-	0.304	0.029	90.7%
	0.05	0.2	49.033	-	-	0.031	89.2%
	-	-	21.551	0.056	0.348	0.037	100%
4.	-	0.2	20.792	0.068	-	0.039	96.5%
(0.023, 0.025, 0.05, 1e6)	0.05	-	21.526	-	0.356	0.038	99.9%
	0.05	0.2	20.647	-	-	0.041	95.8%
	-	-	1.234	0.009	0.277	0.015	100%
5.	-	0.2	1.220	0.009	-	0.015	98.9%
(-0.01, 0.015, 0.7, 1e6)	0.05	-	0.954	-	0.190	0.011	77.4%
	0.05	0.2	0.954	-	-	0.013	77.3%
	-	-	5.586	0.003	0.208	0.025	100%
6.	-	0.2	5.585	0.003	-	0.025	100%
(-0.01, 0.025, 0.7, 1e6)	0.05	-	2.967	-	0.108	0.017	53.1%
	0.05	0.2	2.893	-	-	0.016	51.8%

Table 8: Fixed cost 50,000, RPU 500.

Subplot	Fixed		Results				
$(\mu, \sigma, p_c, n_{max})$	α	β	ExpROI	α	β	δ	ExpROI % of Max
	-	-	0.951	0.237	0.371	0.014	100%
1.	-	0.2	0.925	0.340	-	0.015	97.2%
(0, 0.015, 0.05, 1e6)	0.05	-	0.891	-	0.391	0.023	93.7%
	0.05	0.2	0.876	-	-	0.026	92.1%
	-	-	1.441	0.087	0.372	0.028	100%
2.	-	0.2	1.383	0.108	-	0.030	96.0%
(0, 0.025, 0.05, 1e6)	0.05	-	1.415	-	0.400	0.031	98.2%
	0.05	0.2	1.343	-	-	0.035	93.2%
	-	-	25.800	0.999	0.00	0.00	100%
3.	-	0.2	22.676	0.800	-	0.00	87.9%
(0.023, 0.015, 0.05, 1e6)	0.05	-	17.577	-	0.174	0.028	68.1%
	0.05	0.2	17.557	-	-	0.027	68.1%
	-	-	12.462	0.113	0.216	0.032	100%
4.	-	0.2	12.455	0.116	-	0.032	99.9%
(0.023, 0.025, 0.05, 1e6)	0.05	-	12.060	-	0.238	0.036	96.8%
	0.05	0.2	12.025	-	-	0.037	96.5%
	-	-	0.802	0.014	0.186	0.013	100%
5.	-	0.2	0.802	0.014	-	0.013	100%
(-0.01, 0.015, 0.7, 1e6)	0.05	-	0.721	-	0.157	0.011	89.8%
	0.05	0.2	0.717	-	-	0.010	89.4%
	-	-	3.092	0.005	0.125	0.021	100%
6.	-	0.2	3.044	0.005	-	0.020	98.5%
(-0.01, 0.025, 0.7, 1e6)	0.05	-	2.256	-	0.091	0.016	73.0%
	0.05	0.2	2.173	-	-	0.015	70.3%

Table 9: Fixed cost 100,000, RPU 200.