UPPSALA
UNIVERSITET

SEQUENTIAL A/B TESTING USING PRE-EXPERIMENT DATA

Submitted by
Erik Stenberg

A thesis submitted to the Department of Statistics in partial fulfillment of the requirements for Master degree in Statistics in the Faculty of Social Sciences

Supervisor Patrik Andersson

Spring, 2019

ABSTRACT

This thesis bridges the gap between two popular methods of achieving more efficient online experiments, sequential tests and variance reduction with pre-experiment data. Through simulations, it is shown that there is efficiency to be gained in using control-variates sequentially along with the popular *mixture Sequential Probability Ratio Test*. More efficient tests lead to faster decisions and smaller sample sizes required. The technique proposed is also tested using empirical data on users from the music streaming service Spotify. An R package which includes the main tests applied in this thesis is also presented.

Keywords: A/B testing, sequential analysis, continouous monitoring, variance reduction

CONTENTS

1 Introduction

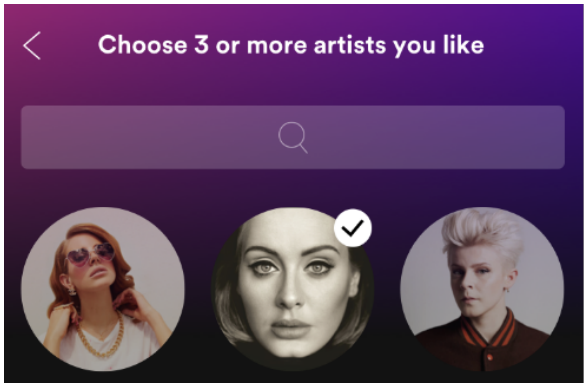
Most major companies present on the web today perform online randomized experiments (ORE's). These experiments are done to evaluate anything from performance of new features, changes in the back-end code to marketing campaigns and promotions. At its essence, a simple ORE with two groups, most commonly referred to as a "split test" or A/B-test, is an experiment based on the idea of exposing users to two different versions, and use a statistical hypothesis test to determine if there is a difference in some variable that we want to test between the two groups. Henceforth, such variables are referred to as *metrics*. This is in line with modern companies' ambition to become increasingly "data-driven". An example of an A/B test at Spotify is illustrated in Figure 1.

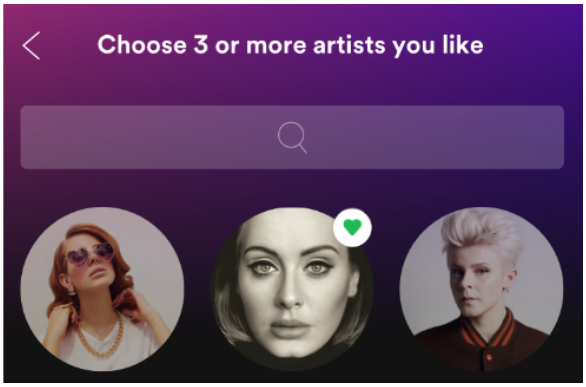
Recent development in technology allow for far better opportunities to store data of various kind, which in turn allows for more experiments, and it is not uncommon for a mature online company to run thousands of experiments every year. Larger tech-companies, such as Netflix, Uber, Spotify and Google have developed advanced experimentation platforms to handle this abundance of experiments. Companies that base their business on experimentation, such as VWO and Optimizely, have also gained ground. Unfortunately, far from all companies running A/B-tests have the experience to avoid stepping into one of a large set of dangerous pitfalls associated with the actual evaluation of an experiment, which Deng, Lu, and Litz (2017) and Dmitriev et al. (2017) discuss. Despite the potential problems of *peeking* at a statistical hypothesis test before required sample size is acquired, this practice is common amongst such companies. In Example 1.1, one such test is illustrated.

1.1 Sequential Analysis

In A/B-tests, a selection of users are randomly allocated to either a treatment group or a reference group (experiments with multiple groups are referred to as A/B/n-tests). The treatment group is exposed to some treatment, while the reference group remains untreated and used for comparison. Throughout this thesis, the metric difference between two such groups will be denoted θ. One major problem involves specifying a minimal detectable effect (MDE), i.e. the smallest θ that the experimenter would like to detect with probability $1 - \beta$. The desired MDE is related to the sample size needed for the test. At large companies, very small effects of treatment are of interest, for example a 0.5 % increase in conversion rate can lead to millions of

FIGURE 1. Example of an A/B test at Spotify. The control group sees the old version, with the tick-mark, and the exposed group sees the new version, with a heart-symbol. When the test is over, it is possible to measure for example how many artists the users clicked.





dollars in increased revenue. Therefore, there is an important trade-off between a small MDE and the potential cost of waiting for more observations. Typically in online experimentation settings, observations arrive to the test in a streaming fashion. For example, if n users are selected to participate in an experiment starting the following day, it is not uncommon to see only a small proportion of those users that actually log in on that day. To avoid disrupting the user experience because of a faulty treatment, a common approach is to keep the sample intake low in the launch period of an experiment, and ramp up sample intake as time goes by. The idea behind this seems intuitive for many practitioners, but can lead to erroneously terminating an experiment before enough individuals are observed to see a statistically significant difference between the groups, and potentially miss an important business opportunity. Companies need fast results, and can sometimes not afford to expose users to an inferior variant during the time a fixed horizon test requires. This makes the standard hypothesis tests less favorable, and the need for testing techniques that allows peeking without inflating Type-I error rates is self-evident.

DEFINITION 1.1 (Sequential Test)**.** *A Sequential Test is a test where the sample size is not pre-determined. Instead, data is collected and evaluated sequentially. For each new observation (or group of observations), a test is performed and a decision is made about whether to continue sampling or stop, according to some pre-defined stopping rule.*

EXAMPLE 1.1. *Let* $X_i \overset{iid}{\sim} N(\theta, 1)$ *with* $i \in \{n, ..., m\}$, $n \geq 1, m > n$ *and we wish to test whether* θ *is equal to* θ_0. *Let* $Z_n = \sqrt{n}(\bar{X}_n - \theta_0)$, *where* $\bar{X}_n$ *is the sample mean after* n *observations are collected. A sequential test that will inflate Type-I error probability well above the nominal significance level* α *is a test where after each new observation* X_{n+1} *we choose to reject* H_0 *if* $|Z_{n+1}| > z_{\alpha/2}$ *or collect a new observation and continue.*

Various techniques that allow for early stopping and sequential analysis of tests have been proposed, many of which were used in clinical trials under the term "interim analysis". See for example Pocock (1977) for Group Sequential Design, Demets and Lan (1994) for the Alpha Spending Function approach, and Bartroff, Shih, and Lai (2013) for a broader overview. More recently, Johari, Pekelis, and Walsh (2015) showed how a method of utilizing a version of the classical Sequential Probability Ratio Test (SPRT) of Wald (1945) will bound the Type-I error rate at nominal level under continuous monitoring of A/B-tests. The test considered by Johari, Pekelis, and Walsh (2015) and many others, called *mixture sequential probability ratio test* first introduced in Robbins (1970) avoids a well-known limitation of the SPRT, namely that a fixed parameter value under H_1 has to be specified.

1.2 VARIANCE REDUCTION

Apart from sequential analysis, techniques to reduce the variance in the metric of interest has gained a lot of attention. Variance reduction is efficient, and simple to implement, and is hence a popular choice by practitioners and researchers aiming to make their tests more sensitive. One such technique is CUPED (Controlled-experiment Using Pre-Experiment Data) and was presented by Deng et al. (2013). The goal was to provide an approach that is generic and applicable to most popular business metrics, such as conversion rates and click-through rates.

On a very high level, the intuition behind CUPED can be described as follows: an experimenter oftentimes has access to pre-experiment data on the same experimental units. These data can be utilized to reduce metric variability and hence improve sensitivity in the test. In fact, the methods applied in CUPED, stratification and control variates, are well-known methods and they have been applied in various Monte Carlo contexts, see Glasserman (2003) for a good overview and detailed explanations. The potential benefits of exploiting this opportunity is easy to see, particularly when there is an abundance of background-variables available on the experimental units, as typically is the case on premium services with paid subscription options.

1.3 Average Treatment Effect

ORE's are generally performed and analyzed under the framework of average treatment effect (ATE) (Xie, Chen, and Shi, 2018). This framework, often referred to as the Rubin causal model (Sekhon, 2008), states that the average treatment effect is the average difference between the users two potential outcomes, namely the one where user i would receive treatment and the one where user i did not receive treatment. This inevitably leads to a missing data problem as only one of the two potential outcomes are observed. With a few assumptions, however, the difference in sample means can be used as an estimator for ATE.

Assumption 1 (Stable unit treatment value assumption, SUTVA)**.**

- *No interference, meaning that units do not interfere with each other. Treatment of one unit does not affect the outcome of another.*
- *Only one treatment version.*

Assumption 2 (Unconfoundedness)**.** *Conditional on a set of covariates, the potential outcomes are independent of treatment assignment.*

It is important to note that these assumptions are subject to violations in many cases. A/B tests are typically performed on websites or in web-applications and we like to think that our treatments are randomly assigned. This is, however, in many cases not entirely true. While it is relatively easy to control many observable covariates, such as country, web-browser, device from which the user is connecting and so on, some covariates are harder. There is one major issue that deserves extra attention. If treatment assignment is randomly assigned upon arrival at the website/application, particularly active users have a considerably larger probability of being assigned treatment than non-active users, as active users have a larger probability of being the next individual who arrives. If the company has access to, for example, a database with registered users and wants to run a test for only registered users, then of course we can randomly assign treatment before arrival to the website/application. This might however be a bad strategy considering that many users are not active at all, and might never actually log in and receive treatment. While this topic is left out from further discussion in this thesis, It is essential to keep this in mind when designing the randomization of treatment. One particular method that suits the scenario of sequential analysis is Sequential Randomization as developed by Zhou et al. (2018).

1.4 Outline

Sequential analysis and variance reduction are mostly based on results that have been extensively used in different areas of statistics, and typically stand as alternatives to each other when moving away from the standard fixed-horizon test with pre-determined sample size. For example, variance reduction is typically applied to an estimator after a fixed-horizon test has been performed to make the test more sensitive (Deng et al., 2013). Theoretically, sequential analysis and certain variance reduction techniques should be orthogonal to each other, and the main contribution of this thesis is to bring these two ideas together to fully utilize the potential win by using variance reduction techniques and the mixture sequential probability ratio test. Ultimately, this will lead to faster decisions, and help companies build more accurate, robust and efficient experimentation platforms. The outline of the rest of this thesis is as follows: first, a discussion on the problem of continuously monitoring t-tests. After that, two popular ways of making A/B tests more efficient are introduced. In section 3, the methodology applied in this thesis is discussed. In section 4 the methodology will be applied to simulations and empirical data from Spotify.

2 Background

2.1 The Peeking Problem

The problem of peeking at experiments is not exclusive to A/B testing, see for example Simmons, Nelson, and Simonsohn (2011), where the authors discuss the amount of false positives that have been published in psychological studies, and shows that this is partly due to researchers' tendency not to wait until experiments are over to analyse the results. The problem, however, becomes a bit more apparent when data arrives in a streaming fashion, as is normally the case for most experimenters on popular online platforms. Experiments and the corresponding p-value processes are normally visualized on dashboard-like interfaces and are easily monitored as new data points arrive to the experiment.

As in many other testing environments, the backbone of A/B-testing is the *t-test*. The t-test requires the experimenter to pre-specify some sample size n as well as Type-I error and Type-II error probabilities. In order to do this, she needs an estimate of the true variance of the variable she sets out to test. Type-I error, or α is usually set to $5\,\%$ or $10\,\%$. Type-II error, or β, and

is usually preferred not to exceed 20 %. This is because we want the test to have a reasonable *Power*. If $\beta = 0.2$ then the power of that test will be , $1 - \beta = 0.8$, that is, 80% probability to reject a false null hypothesis. The t-test ensures that the probability of falsely rejecting a true null hypothesis is α, if, and only if, the experimenter looks at the data one single time, and not decides to do so depending on the data.

The problems with peeking can be argued to be obvious, but it is worth pointing out why and a walk through of a complete example on why intermediate peeking at results will inflate Type-I error is in place. Suppose an experimenter wants to test a new feature on a website. Upon arrival to the website, a visitor is randomly allocated to the treatment group, which is exposed to the new feature, or the reference group that sees the current (unchanged) version of the website. While waiting for visitors to arrive, she continuously looks at the p-value. Let Y denote the treatment group in a metric we intend to test, and the corresponding reference group is denoted by X . Let θ denote the difference between the treated and non-treated in this metric, and $X \sim N(\mu_x, \sigma^2)$, $Y \sim N(\mu_y, \sigma^2)$. Let the true value of θ be zero, i.e. no treatment effect.

PROPOSITION 1. *When drawing samples from a population with finite variance, as $n \to \infty$, a true null hypothesis of zero mean will be rejected with probability one in fixed horizon-testing with continuous monitoring and rejection of H_0 as soon as*

$$|\hat{\theta}_n| > c\sqrt{\hat{\sigma_n}^2/n},$$

where c is some constant, $\hat{\theta}_n$ is our estimate of θ after having collected n observations, and $\hat{\sigma}_n^2/n$ is the estimated variance of $\hat{\theta}_n$. Typically, this c is chosen such that the probability of rejecting a true H_0 without continous monitoring is α.

Proof. *The law of iterated logarithm states that with probability one, there exists arbitrarily large values of n for which*

$$|\hat{\theta}_n| > \sqrt{2 \log \log n/n}.$$

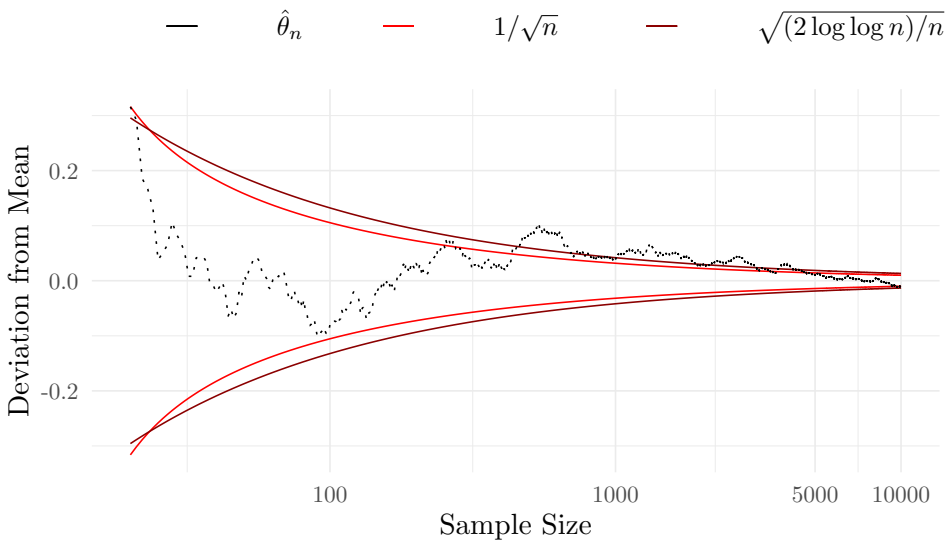
As $n \to \infty$,

$$c\sqrt{\hat{\sigma_n}^2/n} \approx \frac{1}{\sqrt{n}}.$$

Also, $\sqrt{2 \log \log n/n}$ will exceed any threshold proportional to $1/\sqrt{n}$. Hence, $|\hat{\theta}_n|$ will eventually exceed the critical value.

To see the consequences in terms of inflated Type-I error rate when peeking at t-tests and stopping once a significant result is observed, a simulation of 1,000 experiments with 20,000

FIGURE 2. Example of 10,000 iid observations from a standard normal distribution, with the mean computed after each observation (the dotted line). The red line represents a function which, for large enough n, the function represenetd by the dark red line will exceed. Due to the law of iterated logarithm, $|\hat{\theta}|$ will also exceed that function, leading to rejection of the null w.p 1. for large enough n



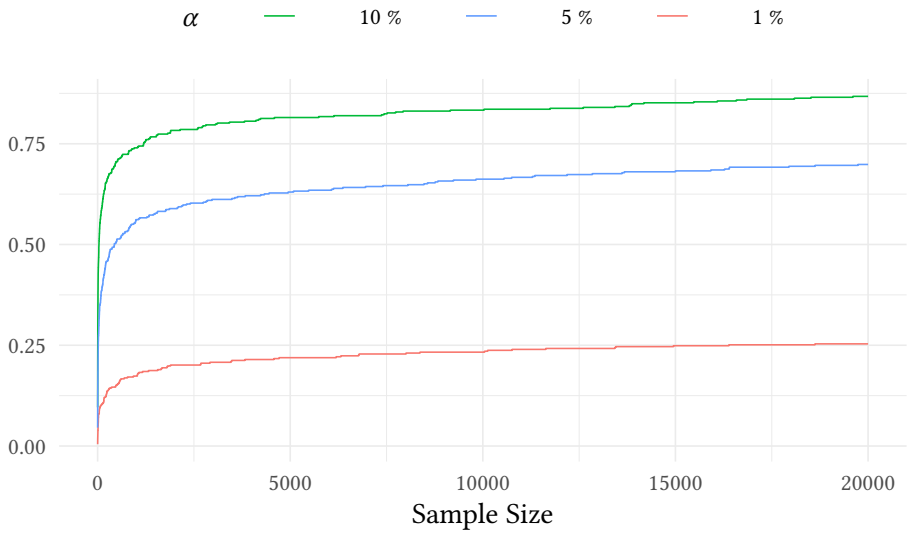
observations each is demonstrated in Figure 3. The tests are run until the p-value is lower than α and then stopped. If α is never reached, the test will run until $n = 20,000$ and H_0 is not rejected. Even with a significance level of 1 %, after 20,000 observations roughly 25 % of the tests have been falsely rejected and a Type-I error has been committed. As implied by Proposition 1, for all significance levels $\alpha \in [0, 1)$, these proportions will tend to 1 if we wait infinitely long.

2.2 THE SEQUENTIAL PROBABILITY RATIO TEST

The Sequential Probability Ratio Test (SPRT) was developed and presented by Abraham Wald (Wald, 1945). The idea is based on the standard likelihood ratio test, which according to the Neyman-Pearson theory of testing simple hypotheses is the most powerful test. Let X be a random variable with a parameter θ, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ be the null and alternative hypothesis respectively. Then the most powerful test of θ which rejects H_0 in favor of H_1 for a significance level α based on N observations is

$$\Lambda = \prod_{i=1}^{N} \frac{f_{\theta_0}(x_i)}{f_{\theta_1}(x_i)} > c,$$

FIGURE 3. Proportion of t-test falsely rejected if rejecting as soon as the p-value is less than α



where c is some constant chosen such that the probability of rejecting H_0 if true is α. For the standard likelihood ratio test approach we pre-specify N and α and find the test which maximizes the probability of rejecting a false null. Wald presented another option: to stop at an intermediate point. While this opens up the opportunity for a test that never ends, this is commonly avoided by setting a *truncation time* where the test is terminated and H_0 accepted. Wald defines a sequential test as any test where you on the basis of the first n observations make one of three decisions: (1) accept H_0, (2) reject H_0 and (3) continue to take a new observation, and consequently treats n as a random variable, and the SPRT after n observations is defined as

$$\Lambda_n = \prod_{i=1}^{n} \frac{f_{\theta_1}(x_i)}{f_{\theta_0}(x_i)}, \tag{1}$$

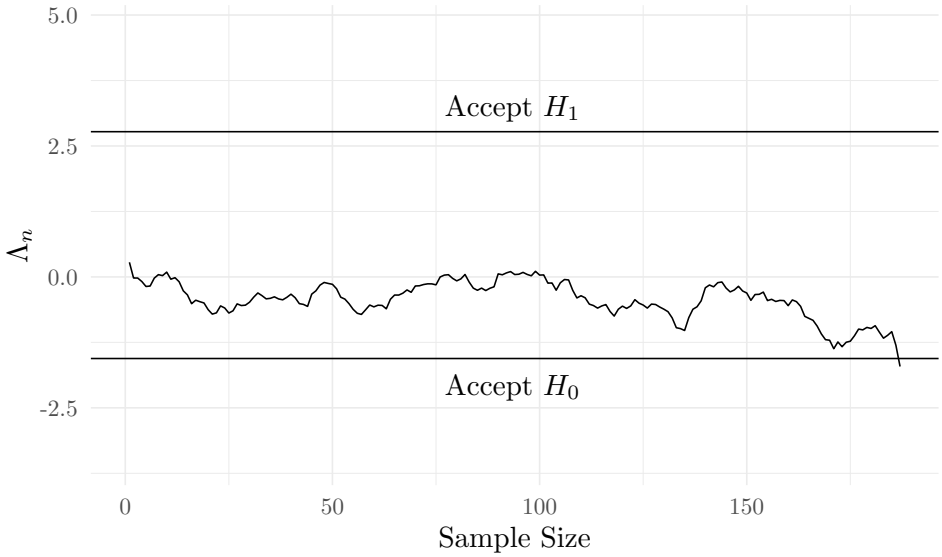
which after each new observation lets the experimenter make one of three decisions: if

- $\Lambda_n < A$, Accept H_0
- $\Lambda_n > B$, Reject H_0 in favor of H_1
- $A < \Lambda_n < B$, collect a new observation and calculate Λ_{n+1},

where $A = \beta/(1-\alpha)$ and $B = (1-\beta)/\alpha$. An example of such a test is demonstrated in Figure 4. If, for any n, $f_{\theta_0}(x_n) = 0$ then we set $\Lambda_n = 0$ without any loss of generality. Typically, the ratio as well as the rejection boundaries are expressed in its respective logarithm. While Wald failed to prove this test *optimal* in his first paper, the SPRT was later proven to be optimal (Wald

and Wolfowitz, 1948) in the sense that it, amongst all sequential tests which do not have larger error probabilities than the SPRT, minimzes the average sample size before a decision is made. Although this approach has turned out successful in many areas, especially when it is important to quickly terminate a harmful experiment, the main limitation is that the experimenter has to specify a constant value of θ under the alternative hypothesis.

FIGURE 4. Example of SPRT of a sample from a standard normal distribution with $H_1 : \theta = 0.1$



2.3 VARIANCE REDUCTION

2.3.1 CONTROL VARIATES

Consider the difference θ between the treatment group Y and the control group X in a variable which we want to perform an A/B test for. Our currently best estimator is $\hat{\theta} = \bar{Y} - \bar{X}$, which satisfies $\mathbb{E}(\hat{\theta}) = \theta$. Let t be another random variable with known expectation τ. We then construct a new estimator

$$\hat{\theta}^* = \hat{\theta} - k(t - \tau), \tag{2}$$

with variance

$$\text{Var}(\hat{\theta}^*) = Var(\hat{\theta}) + k^2\text{Var}(t) - 2\,k\,\text{Cov}(\hat{\theta}, t).$$

If we let

$$k = \frac{\text{Cov}(\hat{\theta}, t)}{\text{Var}(t)},$$

then

$$\begin{aligned}
\text{Var}(\hat{\theta}^*) &= \text{Var}(\hat{\theta}) + \left[\frac{\text{Cov}(\hat{\theta}, t)}{\text{Var}(t)}\right]^2 \text{Var}(t) - 2\frac{[\text{Cov}(\hat{\theta}, t)]^2}{\text{Var}(t)} \\
&= \text{Var}(\hat{\theta}) - \frac{[\text{Cov}(\hat{\theta}, t)]^2}{\text{Var}(t)} \\
&= (1 - \rho^2_{\hat{\theta},t})\,\text{Var}(\hat{\theta}).
\end{aligned}$$

PROPOSITION 2. *The control variate estimator (2) is unbiased, because*

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\hat{\theta} - k(t - \tau)] = \theta - k(\mathbb{E}[\theta] - \tau]) = \theta. \tag{3}$$

By Proposition 2 this new estimator is also an unbiased estimator of θ, but with a variance smaller than $\hat{\theta}$ by a factor of $\rho^2_{\hat{\theta},t}$, and the greater the correlation, the greater the variance reduction. The choice of k above is optimal in the sense that it minimizes the variance of $\hat{\theta}^*$. This method is very common in Monte Carlo sampling, and the obvious difficulty lies in finding another variable t with known expectation which is highly correlated with the metric, but still independent of treatment. Deng et al. (2013) suggested that in online experiments, one of the more efficient ways to go about this problem is to use *pre-experiment* data of the same metric as the control variate.

EXAMPLE 2.1. *Consider an example where we want to reduce the variance of a variable representing the difference between the treatment group, Y and control group, X in a certain metric. Our current estimator is*

$$\hat{\theta} = \bar{Y} - \bar{X},$$

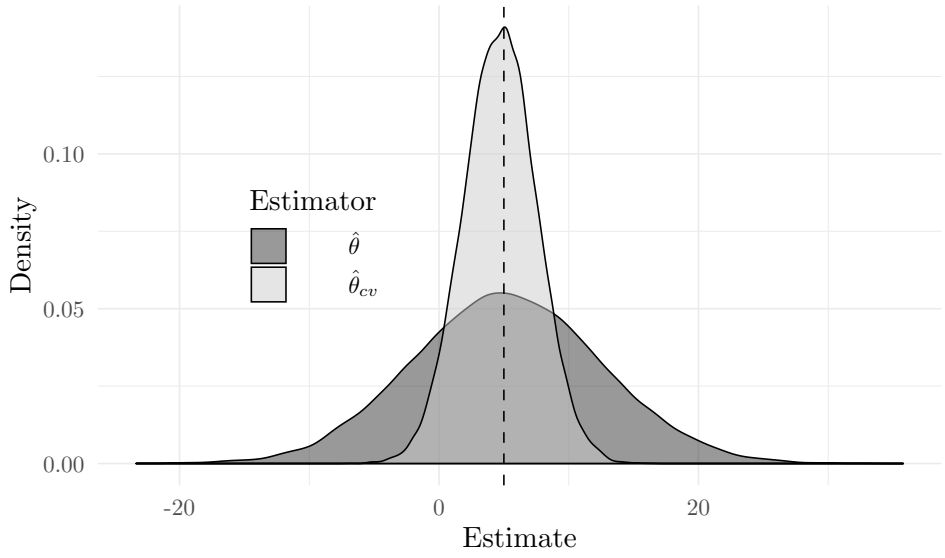
and now let $\bar{Y}_{pre}$ and $\bar{X}_{pre}$ denote the treatment group and control group before launching the experiment. Clearly, $\mathbb{E}[\bar{Y}_{pre} - \bar{X}_{pre}] = 0$ before any treatment has been introduced. Constructing our new estimator $\hat{\theta}_{cv}$ as in (2) yields:

$$\begin{aligned}
\hat{\theta}_{cv} &= \left[\bar{Y} - k\bar{Y}_{pre} + k\,\mathbb{E}[\bar{Y}_{pre}]\right] - \left[\bar{X} - k\bar{X}_{pre} + k\,\mathbb{E}[\bar{X}_{pre}]\right] \\
&= \left[\bar{Y} - k\bar{Y}_{pre}\right] - \left[\bar{X} - k\bar{X}_{pre}\right].
\end{aligned}$$

With the optimal choice of k as above we will have a variance reduction of $(1 - \rho^2)$, where ρ is the correlation between pre-experiment data and post-treatment data. In practice pooled

between the treatment group and control group to ensure unbiasedness. The sampling distributions both $\hat{\theta}_{cv}$ and $\hat{\theta}$ from 30,000 samples of $N = 20{,}000$ from a true A/A test with simulated treatment effect are illustrated in Figure 5.

FIGURE 5. Sampling distribution of estimators after variance reduction using only pre-experiment period data. The parameter estimated is a metric difference from an A/A test previously performed at Spotify. True effect is represented by a dashed line, dark gray for the standard sample mean difference, and light gray for the variance-reduced estimator.



2.3.2 MULTIPLE CONTROL VARIATES

The number of control variates can easily be extended to a higher number. The estimator $\hat{\theta}_{cv}$ still have the same properties of unbiasedness (Proposition 2) and reduced variance as before, only now we hope that we can further reduce the variance. The generalization to l control variates is not more complicated than the extension of standard ordinary least squares with one independent variable to Ordinary Least Squares with p variables. With the above assumptions of equal expectation of the control variates before treatment, the estimator for the difference between the treated group and control group is:

$$\hat{\theta}_{cv} = \left[\bar{Y} - \mathbf{k}^T\bar{\mathbf{Y}}\right] - \left[\bar{X} - \mathbf{k}^T\bar{\mathbf{X}}\right], \tag{4}$$

where now $\mathbf{k}$ is a vector of the parameter estimates of the post-treatment data on a vector of pre-experiment data. i.e.

$$\mathbf{k} = \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}, \tag{5}$$

where $\mathbf{S}$ are the corresponding estimated covariance matrices. The variance is then reduced by a factor of R^2, the coefficient of determination when regressing Y on the covariates $\mathbf{X}$

$$\text{Var}(\hat{\theta}^*) = (1 - R^2)\,\text{Var}(\hat{\theta}).$$

3 Methodology

The main hypothesis test considered in this study is the *mixture Sequential Probability Ratio Test* (henceforth mSPRT), first introduced in Robbins (1970). This, one of many, extensions of Wald's SPRT is particularly useful when experimenters have a large amount of data and want the option of stopping a potentially harmful test at an early stage, or simply obtain results quickly. The mSPRT is relatively new in A/B testing, first applied in Johari, Pekelis, and Walsh (2015).

Before going into the details of the mSPRT it is worth pointing out that if $\{X_i\}_{i=1}^n \sim f_\theta(x_n)$ is a sequence of iid random variables, $f_{\theta_1}(x_n)$ and $f_{\theta_0}(x_n)$ are different probability distributions (and $f_0(x_n) > 0$), then the likelihood ratio $\Lambda_n = f_{\theta_1}(x_n)/f_{\theta_0}(x_n)$ is a martingale under the hypothesis that the common pdf for $\{X_i\}_{i=1}^n$ is $f_{\theta_0}(x_n)$. In the mSPRT, we also let $\pi(\theta) > 0$ denote a mixture (or prior) distribution of the true difference θ between two groups that are to be tested. The motivation behind this is that we almost certainly have an idea about the functional form of the effect of an experiment. Oftentimes, this is much easier than trying to specify a reasonable minimal detectable effect, as required by the SPRT and fixed-horizon tests. Suppose $\theta_0 = 0$ (all members of the exponential family can be centered) and define the mSPRT as:

$$\tilde{\Lambda}_n = \int_{\theta\in\Theta} \Lambda_n\,\pi(\theta)\,d\theta = \int_{\theta\in\Theta} \prod_{i=1}^{n} \frac{f_\theta(x_i)}{f_0(x_i)}\pi(\theta)\,d\theta. \tag{6}$$

That is, we simply integrate the SPRT test statistic over some prior probability space for our parameter under H_1. Since $\tilde{\Lambda}_n$ is a martingale under the null hypothesis, we can use Doob's martingale inequality to conclude that

$$\mathbb{P}_{\theta_0}\left[\tilde{\Lambda}_n > \frac{1}{b}, n \geq 1\right] \leq b \text{ for any } b > 0,$$

and so a natural stopping rule for the mSPRT is

$$\inf\left[n : \tilde{\Lambda}_n < \alpha^{-1}\right]. \tag{7}$$

If we after collecting n observations define the p-value as

$$p_n = \min\left[1, \min(\tilde{\Lambda}_t^{-1} \,:\, t \leq n)\right], \tag{8}$$

then it follows immediately that the probability of the p-value falling below α at any time during the experiment is less than α, i.e.

$$\mathbb{P}_{\theta_0}\left[p_n \leq \alpha\right] \leq \alpha \text{ for any } n. \tag{9}$$

The reader is referred to Johari, Pekelis, and Walsh (2015) for a more elaborate proof of this for all members of the exponential family. For now, let us investigate the application of the mSPRT to the normal distribution.

3.1 Data types and mSPRT

In Johari, Pekelis, and Walsh (2015) and in a companion paper (Johari et al., 2017) it is stated without proof that when using a normal mixing distribution and normal data, it is possible to obtain an explicit formula for the mSPRT. Let $\pi(\theta)$ denote the normal prior probability distribution of θ, the true difference between treated Y and control X, both normally distributed with equal variance σ^2. The closed form for the mSPRT is:

$$\tilde{\Lambda}_n = \int_{\theta\in\Theta} \Lambda_n \, \pi(\theta)\, d\theta = \sqrt{\frac{2\sigma^2}{2\sigma^2 + n\tau^2}} \;\; \exp\left\{\frac{\tau^2 n^2 (\bar{Y}_n - \bar{X}_n - \theta_0)^2}{4\sigma^2(2\sigma^2 + n\tau^2)}\right\}, \tag{10}$$

where τ^2 is the variance of the mixing distribution. Proof of this is provided in Appendix A. Similarly, for binary data, approximate Type-I error control is obtained with small α in tests that require a large n to reject H_0, which typically is the case in A/B-testing settings, using the following mSPRT:

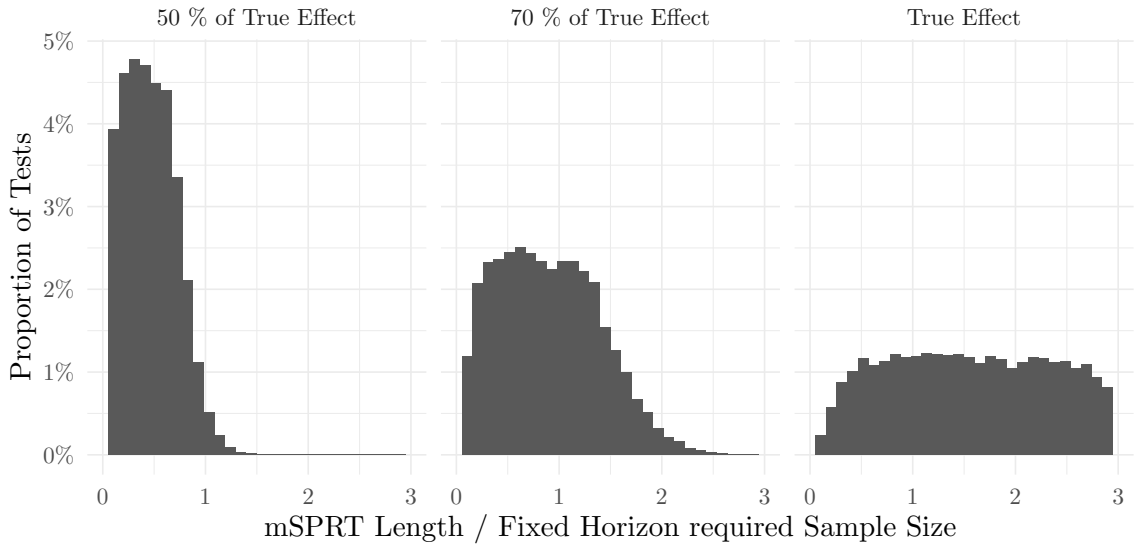
$$\tilde{\Lambda}_n = \int_{\theta\in\Theta} \Lambda_n \, \pi(\theta)\, d\theta = \sqrt{\frac{V_n}{V_n + n\tau^2}} \exp\left\{\frac{n^2\tau^2(\bar{Y}_n - \bar{X}_n - \theta_0)^2}{2V_n(V_n + n\tau^2)}\right\}, \tag{11}$$

where $V_n = \bar{X}_n(1-\bar{X}_n) + \bar{Y}_n(1-\bar{Y}_n)$. Observe that this approximation holds at large n. Applying this to A/B tests the experimenter needs to set the mixing or prior variance τ^2. Mature online companies have already run numerous tests of similar kind, and can quite effectively estimate this variance from previous experiments. It is common to assume a normal mixing distribution because this has frequently been observed in practice and the large-sample distribution of estimated effects tend to be normally distributed. The benefit of assuming a mixture model instead of specifying a simple alternative hypothesis is that these experiments usually

target very small effects, and a misspecified MDE can reduce the power of both t-tests and the SPRT, whereas the mSPRT is robust to misspecified mixture variance

Figure 6 illustrates the run-length of the mSPRT as percentage of what a corresponding fixed-horizon t-test would require in terms of sample size. The simulations are based on results from 20,000 repetitions of tests of samples from a $N(0.05, 0.1)$ representing the difference between treatment group and control group, with known variance. The variance and treatment effect is set such that the realized power of both tests are 80%. That is, the true difference is 0.05, we sample from a $\mathrm{N}(0.05, 0.1)$, run an mSPRT with truncation time $n = 2000$ and compare the number of observations required for the mSPRT to reject the false H_0 to what an the fixed-horizon t-test would require. The closer we set the MDE of the fixed horizon test to the true effect, the less efficiency is gained from using mSPRT, as can be seen in the rightmost plot where the mSPRT required on average 183% of the sample sizes required by fixed-horizon, but only 46 % when the MDE estimation is 50 percent lower than the true effect. It is important to keep in mind that these are common numbers when experimenters are chasing small effects. An MDE of 0.5 % when the actual effect is 1% is 50 % below the true effect.

FIGURE 6. Length of mSPRT as percentage of fixed horizon tests for different estimates of MDE



3.2 CONTROL VARIATES & MSPRT

The main contribution of this thesis is a practical formalization of the joint usage of *control variates* using pre-experiment data and the *mSPRT*, (henceforth *cv-mSPRT*). As discussed in Section 3.1, the closed form of $\tilde{\Lambda}$ at time n with binary data and normal prior is

$$\tilde{\Lambda}_n = \sqrt{\frac{V_n}{V_n + n\tau^2}} \exp\left(\frac{n^2\tau^2(\bar{Y}_n - \bar{X}_n - \theta_0)^2}{2V_n(V_n + n\tau^2)}\right). \tag{12}$$

Now if we follow the strategy outlined in Section 2.3.1, let us first denote the estimator of difference θ between the treatment group and control group after having collected n observations

$$\Delta_n = \bar{Y}_n - \bar{X}_n. \tag{13}$$

We would like to reduce the variance in our estimator by constructing a new estimator $\Delta_n^{(cv)}$. If we choose k as outlined in Section 2.3.1, then a control variate candidate for the true difference is

$$\Delta_n^{(cv)} = (\bar{Y}_n - k\bar{Y}^{(pre)} + k\mathbb{E}\bar{Y}^{(pre)}]) - (\bar{X} - k\bar{X}^{(pre)} + k\mathbb{E}\bar{X}^{(pre)}]), \tag{14}$$

where $\bar{Y}^{(pre)}$ and $\bar{X}^{(pre)}$ is the treatment group and control group means before launching the experiment. At this stage,

$$\mathbb{E}[\bar{Y}^{(pre)}] = \mathbb{E}[\bar{X}^{(pre)}], \tag{15}$$

hence this in an unbiased estimator with expectation

$$\mathbb{E}[\Delta_n^{(cv)}] = \mathbb{E}[\Delta_n] = \mathbb{E}[\bar{Y}_n - \bar{X}_n] = \theta,$$

and variance

$$\text{Var}(\Delta_n^{(cv)}) = (1 - \rho^2)\text{Var}(\Delta_n),$$

as described in Section 2.3.1. Note that the condition in (15) is crucial for $\Delta_n^{(cv)}$ to be unbiased. This is usually not an issue in practice due to the randomization of treatment and control assignment. Following Johari, Pekelis, and Walsh (2015), the asymptotic distribution of Δ_n is $N(\theta, V_n/n)$ where

$$V_n = \bar{X}_n(1 - \bar{X}_n) + \bar{Y}_n(1 - \bar{Y}_n).$$

Hence, our new estimator has asymptotic distribution $N(\theta, V_n(1 - \rho^2)/n)$ and our cv-mSPRT becomes:

$$\tilde{\Lambda}_n^{cv} = \sqrt{\frac{V_n(1 - \rho^2)}{V_n(1 - \rho^2) + n\tau^2}} \exp\left(\frac{n^2\tau^2(\Delta_n^{(cv)} - \theta_0)^2}{2V_n(1 - \rho^2)[V_n(1 - \rho^2) + n\tau^2]}\right). \tag{16}$$

And similarly, for normal data and normal prior,

$$\tilde{\Lambda}_n^{cv} = \sqrt{\frac{2\sigma^2(1-\rho^2)}{2\sigma^2(1-\rho^2)+n\tau^2}} \exp\left(\frac{n^2\tau^2(\Delta_n^{(cv)}-\theta_0)^2}{4\sigma^2(1-\rho^2)[2\sigma^2(1-\rho^2)+n\tau^2]}\right). \tag{17}$$

In practice, we might not know ρ, the true correlation between pre-experiment data and post-treatment data. The easiest way to go about this is to estimate it from data. Since many of these experiments target small effect, large samples are often required, and decisions are seldom made too early for the law of large number to give us a decent estimate of the true correlation. The effect on the test will likely be negligible.

3.2.1 R PACKAGE: MIXTURESPRT

An R package that contains the cv-mSPRT is available at Github. There are methods of calculating the mSPRT and cv-mSPRT with one vector of pre-experiment data in both R and C++. The package also includes plot and print methods, as well as calculation of optimal mixture variance in the spirit of Johari, Pekelis, and Walsh (2015) Section 4.2.

4 RESULTS

In this section, results from simulated and empirical data are presented. Comparisons are made between the standard mSPRT where the tests are performed on the estimator Δ_n and cv-mSPRT where the tests are performed on the estimator $\Delta_n^{(cv)}$. The main goal is to investigate whether the cv-mSPRT can successfully lower the amount of observations required to reject a false H_0 of no difference between two groups by using pre-experiment data. To ensure that neither of the tests does inflate type-1 error rates, A/A tests are also performed for each test setup, i.e. tests without any actual treatment. K

4.1 SIMULATED BINARY DATA

For each test setup below, samples has been drawn iteratively while performing cv-mSPRT and mSPRT respectively. For the tests of binary data, the following approach has been used. Let $Y_i \sim \text{Ber}(p_1)$ be the pre-experiment data of observation i of this variable. We want the mean of the post-treatment variable X_i to be $\mathbb{E}[X_i] = p_2$ for all i. To ensure a correlation of ρ between the pre-experiment data and post-treatment data, and the desired treatment effect, we let the

two transition probabilities, i.e. the probabilities of changing from 0 to 1 or vice versa, q_1 and q_2 be defined as

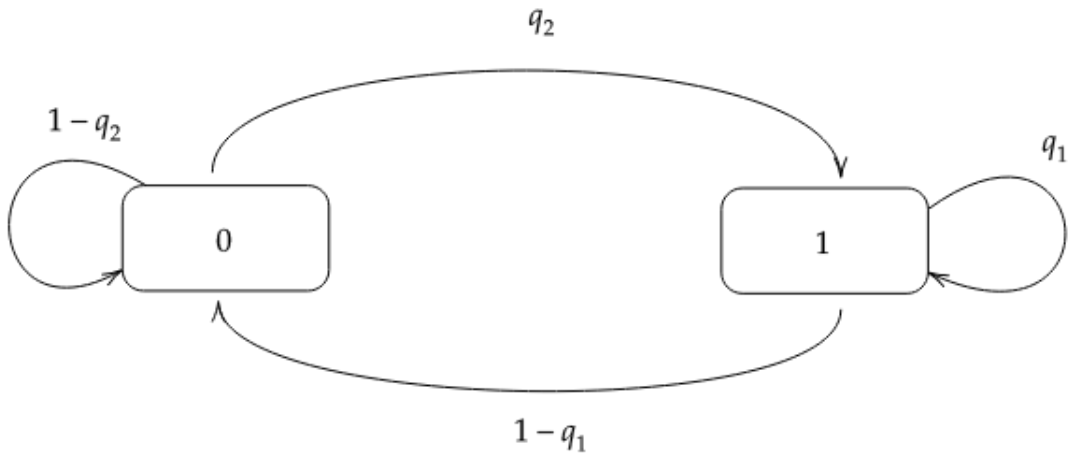
$$\begin{aligned} q_1 &= p_2 + \rho/p_1\sqrt{p_1(1-p_1)p_2(1-p_2)} \\ q_2 &= (p_2 - q_1 p_1)/(1-p_1), \end{aligned} \tag{18}$$

and we construct the post-treatment variable as follows

$$\begin{aligned} P(Y=1|X=1) &= q_1 \\ P(Y=1|X=0) &= q_2, \end{aligned} \tag{19}$$

which will give us correlation between pre-experiment data and post-treatment data of ρ. It will also ensure $\mathbb{E}[Y] = p_1$ as well as $\mathbb{E}[X] = p_2$. In the case of A/A tests, we simply let $p_1 = p_2$. This will induce some transitions, but not change the sample mean in neither group. The rest of the transition probabilities can easily be derived from (18) and (19). The transition scheme is illustrated in Figure 7.

FIGURE 7. Transition Probabilities from pre-experiment to post-treatment with binary data

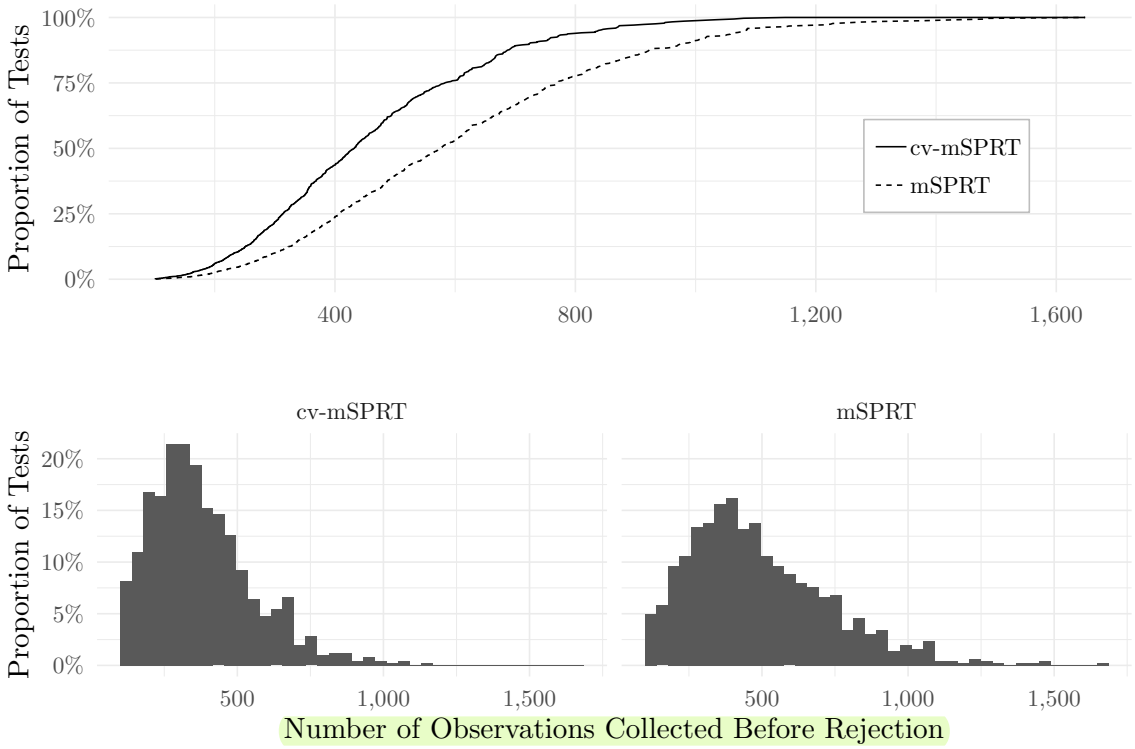


4.1.1 LARGE EFFECT, SMALL SAMPLE

In Figure 8, results from $1,000$ test on binary data are presented. The null hypothesis is, as usual, that there is no difference between the treated and the control. In the bottom part of the figure, we see results from A/B tests in which the post-treatment variable X is generated using $p_1 = 0.5$ and $p_2 = 0.6$, implying a large treatment effect. The upper part of the figure shows cumulative proportion of tests rejected. We note that on average, the cv-mSPRT rejects the

false H_0 after observing only 77 % of what the mSPRT required, while still keeping the type-1 error rate below nominal level.

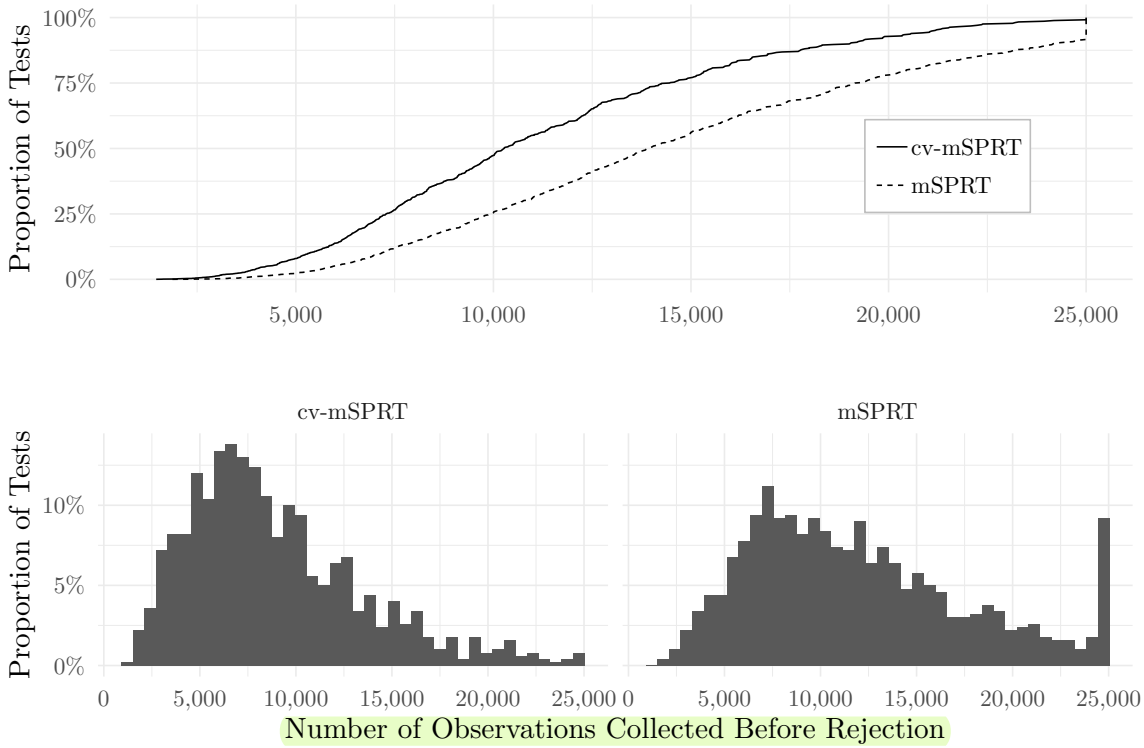
FIGURE 8. Comparison between mSPRT and cv-mSPRT with large effect (10 %) and small sample ($N = 2,000$ for both tests), $\rho = 0.6$. The top plot shows cumulative proportion of tests rejected from $1,000$ tests with binary data. The bottom plot represents the distribution of rejection. Average rejection time for cv-mSPRT is $n = 379$, and for mSPRT $n = 492$



4.1.2 SMALL EFFECT, LARGE SAMPLE

In Figure 9 the results from $1,000$ of each model setup presented below is presented. In this case, the post-treatment variable is generated such that the treatment effect is approximately 2 percentage points, that is, $p_1 = 0.5$ and $p_2 = 0.52$. The average rejection time for cv-mSPRT is approximately 74 % of that of the mSPRT.

FIGURE 9. Comparison between mSPRT and cv-mSPRT with small effect (2 %) and large sample ($N = 25,000$ for each test), $\rho = 0.6$. The bottom plots represent results from $1,000$ tests with binary data. The top plot shows cumulative proportion of tests rejected. Average rejection time for cv-mSPRT is $n = 8,891$, and for mSPRT $n = 11,975$.



4.2 SIMULATED NORMAL DATA

For simulating correlated normal random variables, we may use Choleskys decomposition to simulate our data. We have a desired covariance matrix $\mathbf{\Sigma}$. Let $\mathbf{Z}$ be a vector of random normal variables with zero mean and unit variance. Since a covariance matrix by definition is symmetric and positive definite, $\mathbf{L}\mathbf{L}^T = \mathbf{\Sigma}$. If we then let $\mathbf{X} = \mathbf{L}\mathbf{Z}$ be another vector of random variables, where A is the desired means, then the covariance matrix of $\mathbf{X}$, $\mathbb{E}(\mathbf{X}\mathbf{X}^T)$ can be written as

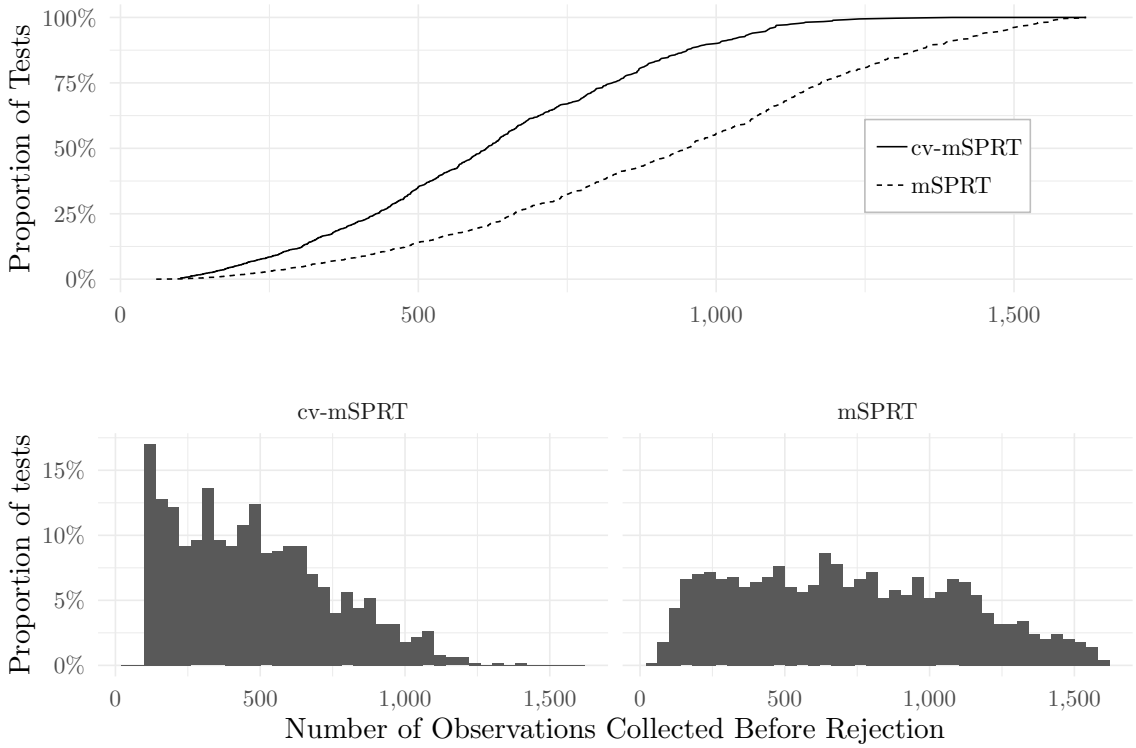
$$\begin{aligned}\mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right) &= \mathbb{E}\left((\mathbf{L}\mathbf{Z})(\mathbf{L}\mathbf{Z})^T\right) \\ &= \mathbb{E}\left(\mathbf{L}\mathbf{Z}\mathbf{Z}^T\mathbf{L}^T\right) \\ &= \mathbf{L}\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^T\right)\mathbf{L}^T \\ &= \mathbf{L}\mathbf{I}\mathbf{L}^T = \mathbf{L}\mathbf{L}^T = \mathbf{\Sigma},\end{aligned} \tag{20}$$

and so if we let $\mathbf{X}$ be our vector of treated and control possibly adding a vector of constants representing the means, we can obtain data with desired means and covariance matrix.

4.2.1 LARGE EFFECT, SMALL SAMPLE

In Figure 10 results from 1,000 tests with truncation time $N = 2,000$ and a 10 % treatment effect of four model setups is presented. The pattern follows that of binary data with the cv-mSPRT requiring on average 66 % of what the regular mSPRT without pre-experiment data requires, while keeping the type-I error rate below nominal level. The correlation between pre-experiment and post-treatment data is here set to 0.6 as well.

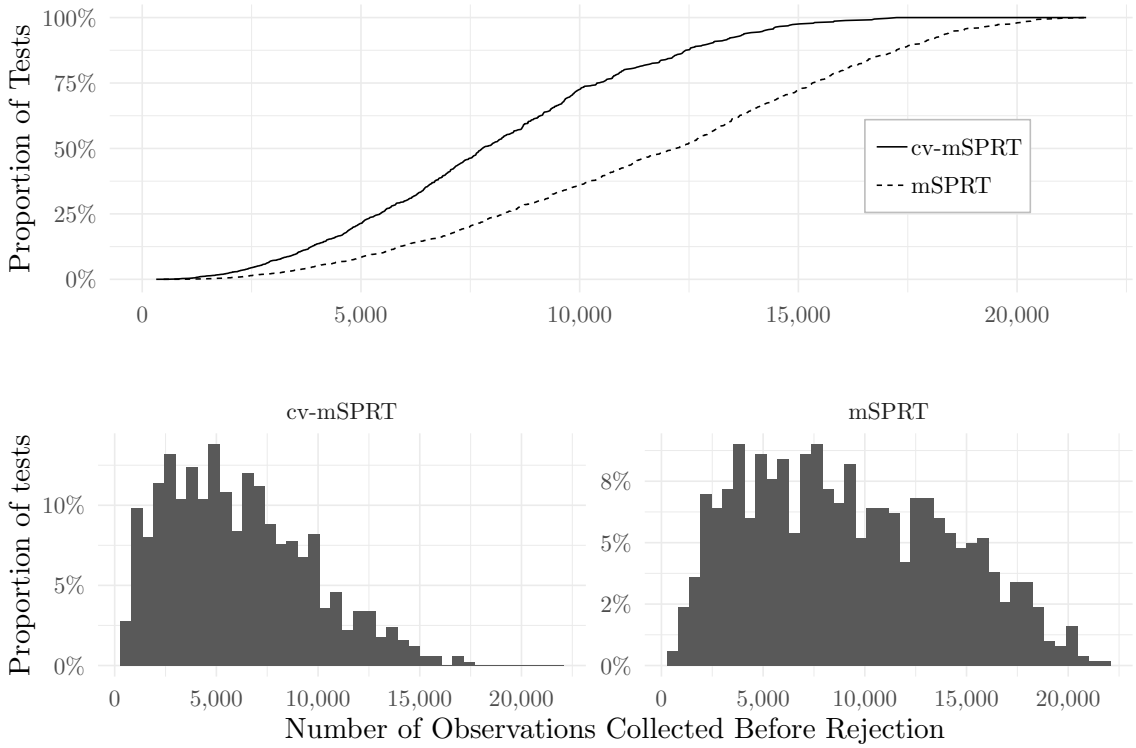
FIGURE 10. Comparison between mSPRT and cv-mSPRT with large effect (10 %) and small sample ($N = 2,000$ for each test), $\rho = 0.6$. The bottom plots represent results from $1,000$ tests with normal data. The top plot shows cumulative proportion of tests rejected. Average rejection time for cv-mSPRT is $n = 477$, and for mSPRT $n = 722$.



4.2.2 SMALL EFFECT, LARGE SAMPLE

Similar to the binary case, with small effect and large sample, we see an average rejection time of around 70 % when using cv-mSPRT of the standard mSPRT when correlation between pre-experiment data and post-experiment data is set to $\rho = 0.6$.

FIGURE 11. Comparison between mSPRT and cv-mSPRT with small effect (2 %) and large sample ($N = 25,000$ for each test), $\rho = 0.6$. The bottom plots represent results from $1,000$ tests with normal data. The top plot shows cumulative proportion of tests rejected. Average rejection time for cv-mSPRT is $n = 6,072$, and for mSPRT $n = 9,211$.



Throughout these simulations, corresponding A/A tests have been conducted to verify that type-I error rates are kept below nominal significance level. The results from the A/B tests above heavily rely on the correlation between pre-experiment data and post-treatment data, as has been discussed previously. A correlation of $\rho = 0.6$ has been used throughout, but the higher correlation, the better variance reduction. Table 1 contains results from the standard mSPRT and cv-mSPRT with three different levels of correlation between pre-experiment data and post-treatment data, and three levels of treatment effect. As expected, when no correlation exists, the cv-mSPRT and mSPRT yield similar results, while an increased correlation yield a better performance for cv-mSPRT. This highlights an important fact that when pre-experiment data is available but correlation is low or non-existing, using cv-mSPRT has negligible effect on the average rejection time for the tests

TABLE 1. Average number of observations required before rejection of (cv-)mSPRT with different treatment effects and correlations between pre-experiment data and post-treatment. Test on normally distributed data with $\mu = 10$ without treatment and $\sigma = 2$. Based on 2,000 tests of each setup.

	Treatment Effect		
Correlation	1 %	2.5 %	5 %
w/o pre-exp. data (mSPRT)	3975	1092	332
$\boldsymbol{\rho = 0}$ (cv-mSPRT)	3991 (100.4 %)	1093 (100.09 %)	336 (101.2 %)
$\boldsymbol{\rho = 0.5}$ (cv-mSPRT)	3916 (98.52 %)	1042 (95.42 %)	316 (95.18 %)
$\boldsymbol{\rho = 0.9}$ (cv-mSPRT)	3665 (92.2 %)	868 (79.49 %)	268 (80.72 %)

4.3 SPOTIFY USER DATA

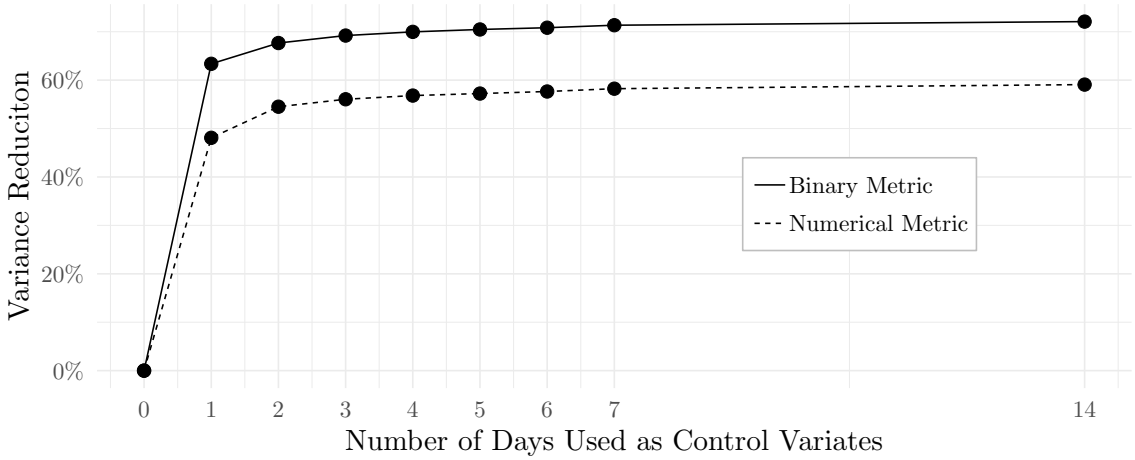
4.3.1 CONTROL VARIATES

As of May 2019, Spotify has approximately 217 million monthly active users. The methods applied so far in the simulations are the mixture Sequential Probability Ratio test and variance reduction with control variates using pre-experiment data. As mentioned before, these two methods have independently been demonstrated to work well in such testing environments with a large number of users arriving in a streaming fashion to the experiment. We have thus far concluded that the efficiency gained from using control variates sequentially highly depend on the correlation between pre-experiment data and post-treatment data. To see what this looks like in practice, a random snapshot of approximately 1,500,000 users has been taken from Spotify's user base. Each user is followed for two weeks time noting two metrics, one binary and one numerical.

From the user snapshot, $20,000$ users are randomly selected. Half of them are assigned to a treatment group and the rest to the control group. Variance reduction using control variates are performed once at the entire sample of $20,000$ users, and finally the estimated difference between the groups are calculated. This was repeated $10,000$ times for each number of days used as control variates to see comparable sampling variances. In Figure 12, the results are demonstrated. Note that the treatment group were not actually exposed to any treatment, i.e. these are $10,000$ A/A tests. So the estimator is simply an estimator of 0.

It is clear that a large part of variance is reduced using only the past day while subsequent days carry small, yet important, variance reduction potential.

FIGURE 12. Variance reduction in metric difference estimator for different number of days used as control variates. The tick at 0 on the x-axis represent the scenario where we don't do any variance reduction at all.



With only four days of control variates, we manage to reduce the variance in the difference estimator by approximately 70 % for the binary metric, and a bit short of 60 % for the numerical metric. This shows how efficient control variates can be on user data. An important clarification to make is that treatment effects in these types of experiments are usually very small, however not irrelevant. Another version of this experiment where a simulated treatment effect in percentage from a $U(0, 0.02)$ (i.e. a treatment effect of 0-2%) for each iteration were performed, and the variance reductions were a few percentage points lower than those presented above. Nonetheless, Figure 12 serves as an indicator of approximately how much variance reduction can be expected from using control variates on these two metrics.

4.3.2 CV-MSPRT ON SPOTIFY USER DATA

Figure 13 and 14 shows the cv-mSPRT on Spotify user data. In total, six different set ups of experiments are explored. For the binary metric and the numeric metric, the cv-mSPRT is tested using one, two and six days of pre-experiment data as control variates. For each setup, one thousand experiments are run and evaluated. The experiments are based on random samples from Spotify's user data. The treatment, however, is simulated. In the binary case, the population from which we sample has a proportion of around 0.5, and treatment is assigned

such that the proportion in the treatment groups are on average 2 percentage points higher than those in the control groups. For the numerical metric, a simulated treatment effect of 10% is added. These effects are not to be thought of as realistic in any sense, rather this is more of a 'proof of concept' implementation of the test to demonstrate that control-variates with pre-experiment data can reduce the number of observations before a correct rejection.

FIGURE 13. cv-mSPRT on Spotify user data on a numerical metric. The y-axis represents the proportion of tests rejected, and the x-axis shows the sample size. The dashed line represents the case when we include the previous six days as control variates, the dotted line when we include the two previous day, and the solid line when we only include one day. The results are based on $1,000$ experiments with simulated treatment effect, and are truncated after $10,000$ observations.

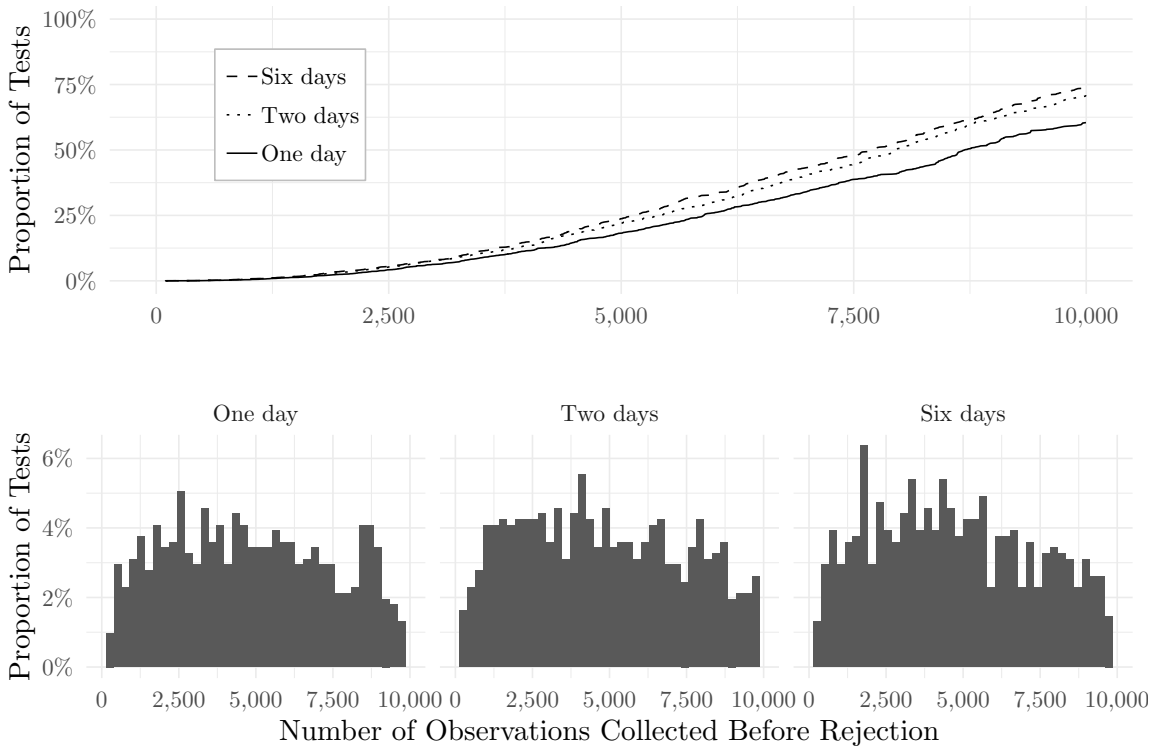
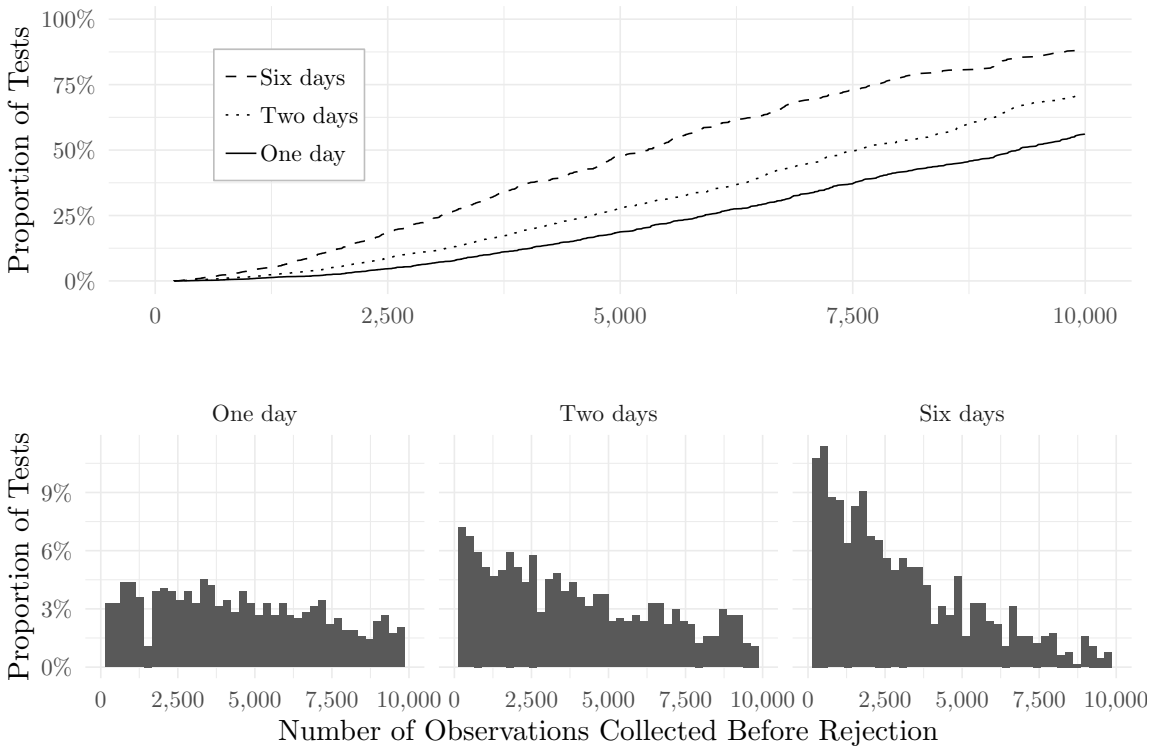


FIGURE 14. cv-mSPRT on Spotify user data on a binary metric. The y-axis represents the proportion of tests rejected, and the x-axis shows the sample size. The dashed line represents the case when we include the previous six days as control variates, the dotted line when we include the two previous day, and the solid line when we only include one day. The results are based on $1,000$ experiments with simulated treatment effect, and are truncated after $10,000$ observations.



5 DISCUSSION & FURTHER WORK

Sequential Analysis in A/B testing is becoming increasingly popular, as it is in line with large companies' ambition to efficiently run experiments and obtain results quickly. One such technique which has been discussed is mixture Sequential Probability Ratio Test. Variance Reduction techniques, such as control variates using pre-experiment data has been applied in fixed-horizon A/B tests to successfully reduce variance and thus reduce the number of observations required to run an experiment. In this thesis, it is demonstrated that these two techniques can be used simultaneously (cv-mSPRT) to further increase efficiency in A/B tests. Reducing the sample size needed for an A/B test has several advantages. It is faster, cheaper and allows for more non-overlapping experiments targeting the same sub-population.

The efficiency of using control variates with pre-experiment data and mSPRT jointly heavily relies on the correlation between pre-experiment data and post-treatment data. One of the important results demonstrated is that, while cv-mSPRT might require additional computational power, there is no efficiency lost if no correlation exists. Using cv-mSPRT when no correlation exists will result in approximately the same test as the standard mSPRT. Higher correlation, however, will yield shorter tests if treatment effect exist.

Using pre-experiment data of more than one day shows some efficiency gain in variance reduction at Spotify user data, but mustn't necessarily be the case elsewhere. A pilot study to estimate these correlations might be worthwhile in this scenario.

While cv-mSPRT at this early stage looks promising, there are still topics that has been left out entirely from discussion. One such topic is heterogeneous treatment effect. Many treatments introduced in A/B tests affect users with, for instance, different behavior in different ways. Even if treatment is randomized, it is reasonable to believe that particularly active users will be exposed more often since they log in to the platform more often. Since sequential tests might not observe all users that were assigned treatment (we might reject early), there is a substantial risk that an unproportionally large part of the exposed users are in fact active users. Active users might not be effected the same way as less active users by the particular treatment, hence introducing some bias. While active versus less active users might be the sole most important such factor, many other potential strata of users might introduce bias, for example different devices, different browsers, country and age. Hence, a natural next extension of this is to put this test in a proper potential outcome framework.

Control variates with pre-experiment data of the same metric as the one tested is the only variance reduction technique discussed here. Another potential extension that is of interest for companies with a much data on their users is to try some covariate selection algorithm to further reduce the variance with help of pre-experiment data of the same metric as well as, for example, transformations of numerous other co-variates available to the experimenter. This could be done by learning a neural net to select those combinations of co-variates that explain the largest part of the naturally occuring variance in the metric of interest.

6 ACKNOWLEDGMENT

I would like to thank my supervisor Patrik Andersson for his support throughout writing this thesis, Steven Corroy at Spotify for putting time and interest in my writing, always giving me insightful comments and actionable feedback. A special thanks to Mattias Frånberg, also at Spotify, for his invaluable help and support.

REFERENCES

Bartroff, Jay, Mei-Chiung Shih, and Tze L. Lai (2013). *Sequential Experimentation in Clinical Trials*.

Demets, David L. and K. K. Gordon Lan (1994). "Interim analysis: The alpha spending function approach". *Statistics in Medicine* 13.13-14, pp. 1341–1352.

Deng, Alex, Jiannan Lu, and Jonthan Litz (2017). "Trustworthy Analysis of Online A/B Tests: Pitfalls, challenges and solutions". English. ACM, pp. 641–649.

Deng, Alex et al. (2013). "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data". English. ACM, pp. 123–132.

Dmitriev, Pavel et al. (2017). "A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments". English. Vol. 129685. ACM, pp. 1427–1436.

Glasserman, Paul (2003). *Monte Carlo Methods in Financial Engineering*. English. Vol. 53. New York, NY: Springer New York.

Johari, Ramesh, Leo Pekelis, and David Walsh (2015). "Always valid inference: Bringing sequential analysis to A/B testing".

Johari, Ramesh et al. (2017). "Peeking at A/B Tests: Why It Matters, and What to Do About It". KDD '17. Halifax, NS, Canada: ACM, pp. 1517–1525.

Pocock, Stuart J. (1977). "Group Sequential Methods in the Design and Analysis of Clinical Trials". English. *Biometrika* 64.2, pp. 191–199.

Robbins, Herbert (1970). "Statistical Methods Related to the Law of the Iterated Logarithm". English. *The Annals of Mathematical Statistics* 41.5, pp. 1397–1409.

Sekhon, Jasjeet S (2008). "The Neyman-Rubin model of causal inference and estimation via matching methods". *The Oxford handbook of political methodology* 2, 1–citation_lastpage.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". English. *Psychological Science* 22.11, pp. 1359–1366.

Wald, A. (1945). "Sequential Tests of Statistical Hypotheses". English. *The Annals of Mathematical Statistics* 16.2, pp. 117–186.

Wald, A. and J. Wolfowitz (1948). "Optimum Character of the Sequential Probability Ratio Test". English. *The Annals of Mathematical Statistics* 19.3, pp. 326–339.

Xie, Yuxiang, Nanyu Chen, and Xiaolin Shi (2018). “False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments”. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, pp. 876–885.

Zhou, Quan et al. (2018). “Sequential rerandomization”. English. *Biometrika* 105.3, pp. 745–752.

A PROOFS

With normal data and normal prior, the mSPRT can be solved analytically:

$$\tilde{\Lambda}_n = \sqrt{\frac{2\sigma^2}{2\sigma^2 + n\tau^2}} \;\; \exp\left\{\frac{\tau^2 n^2(\bar{z} - \theta_0)^2}{4\sigma^2(2\sigma^2 + n\tau^2)}\right\} \tag{21}$$

Proof. Let $X_n = \{X_1, X_2, \ldots, X_n\}$ and $Y_n = \{Y_1, Y_2, \ldots, Y_n\}$ be observations drawn independently from two normally distributed populations, with means μ_Y and μ_X and equal known variance σ^2. Let $Z_n = Y_n - X_n$ such that $Z_n \sim N(\theta, 2\sigma^2)$. We want to test the hypotheses:

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta = \theta_1. \tag{22}$$

Let $\pi(\theta)$ denote the prior density of θ, and let it be $N(\theta_0, \tau^2)$. The mixture probability ratio is defined as:

$$\begin{aligned}
\tilde{\Lambda}_n &= \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} \frac{f_\theta(z_i)}{f_{\theta_0}(z_i)}\right) h(\theta)\; d\theta = \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} \frac{\sqrt{2\pi 2\sigma^2}\exp\left\{-\frac{(z_i-\theta)^2}{4\sigma^2}\right\}}{\sqrt{2\pi 2\sigma^2}\exp\left\{-\frac{(z_i-\theta_0)^2}{4\sigma^2}\right\}}\right) \frac{\exp\left\{-\frac{(\theta-\theta_0)^2}{2\tau^2}\right\}}{\sqrt{2\pi\tau^2}}\; d\theta \\
&= \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} \frac{\exp\left\{-\frac{(z_i-\theta)^2}{4\sigma^2}\right\}}{\exp\left\{-\frac{(z_i-\theta_0)^2}{4\sigma^2}\right\}}\right) \frac{\exp\left\{-\frac{(\theta-\theta_0)^2}{2\tau^2}\right\}}{\sqrt{2\pi\tau^2}}\; d\theta \\
&= \frac{\frac{1}{\sqrt{2\pi\tau^2}}}{\prod_{i=1}^{n}\exp\left\{-\frac{1}{4\sigma^2}(z_i-\theta_0)\right\}} \int_{\infty}^{\infty} \exp\left\{-\frac{1}{2\tau^2}(\theta_0-\theta)^2\right\} \prod_{i=1}^{n} \exp\left\{-\frac{1}{4\sigma^2}(z_i-\theta)^2\right\}\; d\theta.
\end{aligned}$$

Continuing only with the integral part of the equation

$$\begin{aligned}
&\int_{\infty}^{\infty} \exp\left\{-\frac{1}{2\tau^2}(\theta_0-\theta)^2\right\} \prod_{i=1}^{n} \exp\left\{-\frac{1}{4\sigma^2}(z_i-\theta)^2\right\}\; d\theta = \\
&\int_{\infty}^{\infty} \exp\left\{\theta^2\left[-\frac{1}{2\tau^2} - \frac{n}{4\sigma^2}\right] - \theta_0^2\left[\frac{1}{2\tau^2}\right] + \theta\left[\frac{2\theta_0}{2\tau^2} + \frac{2\sum_{i=1}^{n} z_i}{4\sigma^2}\right] - \frac{\sum_{i=1}^{n} z_i^2}{4\sigma^2}\right\}\; d\theta
\end{aligned}$$

by algebra. And solving this integral using standard formulas yields

$$\frac{\sqrt{\pi}}{\sqrt{\frac{1}{2\tau^2} + \frac{n}{4\sigma^2}}} \;\; \exp\left\{(\frac{\theta_0}{2\tau^2} + \frac{\sum_{i=1}^{n} z_i^2}{4\sigma^2})/\left(\frac{1}{2\tau^2} + \frac{n}{4\sigma^2}\right)\right\}.$$

Putting this back into the original equation and rearranging a bit and use a little algebra to cancel terms on the factors separately gives us

$$\frac{\frac{\sqrt{\pi}}{\sqrt{2\pi\tau^2}}}{\sqrt{\frac{1}{2\tau^2} + \frac{n}{4\sigma^2}}} \;\; \frac{\exp\left\{\frac{\frac{\theta_0}{2\tau^2} + \frac{\sum_{i=1}^{n} z_i^2}{4\sigma^2}}{\left(\frac{1}{2\tau^2} + \frac{n}{4\sigma^2}\right)}\right\}}{\prod_{i=1}^{n}\exp\left\{-\frac{1}{4\sigma^2}(z_i-\theta_0)^2\right\}} = \sqrt{\frac{2\sigma^2}{2\sigma^2 + n\tau^2}} \;\; \exp\left\{\frac{\tau^2 n^2(\bar{z} - \theta_0)^2}{4\sigma^2(2\sigma^2 + n\tau^2)}\right\} \quad \square$$

B R PACKAGE: 'MIXTURESPRT'

Package 'mixtureSPRT'

May 23, 2019

Type Package

Title Mixture Sequential Probability Ratio Test

Version 1.0

Date 2019-04-02

Author Erik Stenberg

Maintainer Erik Stenberg <erik.stnb@gmail.com>

Description Perfoms mixture Sequential Probability Ratio Test for normally and Bernoulli distributed data, with methods in R and C++.

License GPL (>= 2)

Imports Rcpp,
ggplot2

LinkingTo Rcpp

RoxygenNote 6.1.1

Suggests knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

H_0 Null hypothesis of statistical test. 10, 11, 16, 17

H_1 Alternative hypothesis of statistical test. 6, 10, 11

Fixed-horizon test A statistical test that requires a pre-specified sampe size. 35

Power Probability to correctly reject a false null hypothesis. 9

Type-I error Rejection of a true null hypothesis. The probability of Type-I error is here denoted α. 5, 6, 8–10, 16, 35

Type-II error Failure to reject a false null hypothesis. The probability of Type-II error is here denoted β. 8