A Sequential Test for Selecting the Better Variant

Online A/B testing, Adaptive Allocation, and Continuous Monitoring

Nianqiao Ju
Harvard University
Cambridge, MA
nju@g.harvard.edu

Diane Hu
Etsy Inc.
Brooklyn, NY
dhu@etsy.com

Adam Henderson
Etsy Inc.
Brooklyn, NY
ahenderson@etsy.com

Liangjie Hong
Etsy Inc.
Brooklyn, NY
lhong@etsy.com

ABSTRACT

Online A/B tests play an instrumental role for Internet companies to improve products and technologies in a data-driven manner. An online A/B test, in its most straightforward form, can be treated as a static hypothesis test where traditional statistical tools such as p-values and power analysis might be applied to help decision makers determine which variant performs better. However, a static A/B test presents both time cost and the opportunity cost for rapid product iterations. For time cost, a fast-paced product evolution pushes its shareholders to consistently monitor results from online A/B experiments, which usually invites peeking and altering experimental designs as data collected. It is recognized that this flexibility might harm statistical guarantees if not introduced in the right way, especially when online tests are considered as static hypothesis tests. For opportunity cost, a static test usually entails a static allocation of users into different variants, which prevents an immediate roll-out of the better version to larger audience or risks of alienating users who may suffer from a bad experience. While some works try to tackle these challenges, no prior method focuses on a holistic solution to both issues.

In this paper, we propose a unified framework utilizing sequential analysis and multi-armed bandit to address time cost and the opportunity cost of static online tests simultaneously. In particular, we present an imputed sequential Girshick test that accommodates online data and dynamic allocation of data. The unobserved potential outcomes are treated as missing data and are imputed using empirical averages. Focusing on the binomial model, we demonstrate that the proposed imputed Girshick test achieves Type-I error and power control with both a fixed allocation ratio and an adaptive allocation such as Thompson Sampling through extensive experiments. In addition, we also run experiments on historical Etsy.com A/B tests to show the reduction in opportunity cost when using the proposed method.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5940-5/19/02.
https://doi.org/10.1145/3289600.3291025

CCS CONCEPTS

• **Mathematics of computing** → **Probabilistic inference problems**; • **General and reference** → *Experimentation*; • **Information systems** → Online analytical processing;

KEYWORDS

Online A/B Tests, Controlled Experiments, Sequential Analysis, Imputed Sequential Analysis, Thompson Sampling

ACM Reference Format:
Nianqiao Ju, Diane Hu, Adam Henderson, and Liangjie Hong. 2019. A Sequential Test for Selecting the Better Variant: Online A/B testing, Adaptive Allocation, and Continuous Monitoring. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291025

1 INTRODUCTION

In the current landscape of Internet companies, rapid iteration is the key to product success. Online A/B testing[1], as the *de-facto* method for such process, provides a scientific way for comparing multiple variants and ultimately choosing the one that improves a company-aligned metric with greatest confidence. An online A/B experiment, in its most straightforward form, can be considered as an online *static* A/B testing following a Null Hypothesis Statistical Testing (NHST) framework. While it provides traditional statistical tools such as p-values and power analysis to help decision makers determine which variant performs better, NHST also imposes a number of requirements on the experimentation setup that may make an experiment infeasible or costly in practice. In general, two types of costs exist in NHST framework, namely *time cost* and *opporunity cost*, placing inherent challenges to fast product iterations.

For time cost, the main issue is that an experiment following NHST requires a fixed sample size and therefore a fixed time window, which does not allow repeated significant testing, or "continuous monitoring" [3, 6]. In particular, under such a framework, an experiment sample size needs to be fixed in advance, as a function of effect size, power, and significance levels. This may lead to a pre-determined sample size that is prohibitively large, especially when the difference to be detected is small (as is often the case with challenging business metrics, such as conversion rate, that are

[1] In this paper, we use "testing", "test" and "experiment" interchangeably.

difficult to move). In such a scenario, it may take months or even years to collect enough samples in order to properly power and conclude a test. As mentioned, the NHST framework also prohibits continuous monitoring. This means that statistical guarantees only hold once all samples, up to a pre-determined sample size, have been observed. In practice, this "no peeking" rule can be difficult to follow, especially when early results look compelling. Pressure from stakeholders to improve metrics and iterate quickly often lead practitioners to draw conclusions about an experiment prematurely, which violates the assumption of a fixed sample size, and ultimately renders the result of the A/B test invalid.

For opportunity cost, it primarily comes from a *static allocation* of users for the duration of an experiment, which is imposed by the NHST framework. In a typical static A/B test setting, users are pre-allocated into the control or treatment group in advance. The ratio of users allocated to each variant (e.g., 50% v.s. 50%) is also fixed throughout the experiment. This means that no matter how good or bad a user experience is in either variant, they must stay in that variant for the remainder of the experiment period. If one variant outperforms the other early on in the experiment, a static allocation setting prevents an immediate roll-out of the better version until the conclusion. While this may be acceptable in most cases, this method runs the risk of alienating users who may suffer from a bad experience in the sub-optimal variant.

To address time cost, sequential testing (ST) as a principled methodology has been developed in statistics. In a nutshell, ST allows intermediate checks of significance, enabling continuous monitoring. In addition, ST can help decision makers conclude an experiment earlier with often much fewer samples than the pre-determined sample size. Because of these attractive properties, ST is becoming more widely adopted in Internet companies that seek for more rapid product evolutions. The most commonly used method of ST follows the Sequential Probability Ratio Test (SPRT) proposed by Wald [20], providing a frequentist-view of the *optional stopping criterion* for A/B tests. Later, Robbins [13] developed a class of STs named mixture Sequential Probability Ratio Test (mSPRT), which is widely used by researchers and practitioners. It is further utilized by Johari et al. [5, 6] to establish an *always valid p*-value for online experiments to mitigate the problem of *p*-value peeking. However, the main drawback of this line of work is that they primarily focus on the hypothesis setting $H_0 : \theta_1 = \theta_2$ vs $H_1 : \theta_1 \neq \theta_2$, which cannot be used to distinguish which variant is truly better. In addition, a controversy of mSPRT is that it is a test of *power one* [14], meaning that under any alternative $\theta_1 \neq \theta_2$, the test is eventually guaranteed to reject the null hypothesis, if the user is willing to wait long enough. This implies that the test potentially never terminates, leading to a sub-optimal situation for online tests where rapid decision-making is required.

To address opportunity cost, Scott from Google [17] discusses how Google utilizes Multi-Armed Bandit (MAB) techniques to develop an adaptive scheme for allocation on their platform. This technique uses a stopping rule based on percentage reward, which is different from the cumulative regret studied in the MAB literature. In addition, Johari et al. [6] discusses the effect of using MAB inside mSPRT but does not provide any experimental results.

In this paper, we provide a unified framework to address these two major costs from online static A/B tests *simultaneously*. In particular, we develop a novel ST such that the better variant can be identified, offering more effective decision-making than simply detecting the difference. Additionally, the proposed approach exploits Thompson sampling, a specific form of MAB, to achieve an adaptive allocation for users in an experiment. The proposed method optimizes the outcome (reward) of the experiment while reducing the risk of losing users who are exposed to the sub-optimal variant. Our contributions are as follows:

(1) Propose an imputed sequential Girshick test for Bernoulli model with a fixed allocation.
(2) Use simulations to demonstrate that the test procedure also applies to an adaptive allocation such as Thompson sampling with a small error inflation.
(3) Conduct a regret analysis of A/B tests from the MAB perspective.
(4) Conduct extensive studies including simulations as well as experiments on an industry-scale experiment, demonstrating the effectiveness of the proposed method and offering practical considerations.

The remainder of this paper is structured as follows: In Section 2, we introduce related work. In Section 3, we introduce the proposed framework. We conduct simulations in Section 4, and then show results from industrial-scale experiments in Section 5. Finally, in Section 6, we conclude the paper and present future directions.

2 RELATED WORKS

In this section, we provide some preliminaries and related works in three of the areas that our proposed method touches on.

2.1 Sequential Probability Ratio Test

Proposed by Abraham Wald, sequential analysis [20] studies experiments where the number of observations required is not determined in advance and at each stage of the experiment a decision is made to accept some hypothesis, reject it, or take more observations. In this section we review the sequential probability ratio test (SPRT) for testing simple hypotheses.

Consider a one-variant experiment concerning a random variable $X \sim f_\theta(\cdot)$ where $\theta \in \Theta \subset \mathbb{R}$ and with two simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ (assuming $\theta_0 < \theta_1$ without loss of generality). Because we want to make decisions based on the true value of θ, and based on our risk tolerance, we should divide the domain of θ into three parts: $\omega_a = \{\theta : \theta < \theta_0\}$ region with preference for acceptance of H_0 , $\omega_r = \{\theta : \theta > \theta_1\}$ region with preference for rejection of H_0, and a region of indifference.

The sequential probability ratio test has strength (α, β) in the sense that

$$\text{Type-I error} \quad = \mathbb{P}(\text{reject } H_0 | \theta < \theta_0) \quad \leq \alpha \tag{1}$$

$$\text{Power} \quad = \mathbb{P}(\text{reject } H_0 | \theta > \theta_1) \quad \geq 1 - \beta. \tag{2}$$

For a given choice of constants $A, B \geq 0$, at each stage of the experiment we compute the probability ratio

$$\frac{p_{1m}}{p_{0m}} = \frac{f_{\theta_1}(x_{1:m})}{f_{\theta_0}(x_{1:m})}.$$

We continue the experiment and take more observations if $B < \frac{p_{1m}}{p_{0m}} < A$; if $\frac{p_{1m}}{p_{0m}} > A$, then the process terminates with a decision to reject H_0; and if $\frac{p_{1m}}{p_{0m}} < B$ then we termiante with acceptance of H_0. The choice of A, B determines the strength (α, β) - with a practical approximation of this relationship given by $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$. This choice of A, B might inflate total errors, but Wald has shown that at most one of the Type I or II error would increase [20].

Although the SPRT finishes in finite time with probability 1, an experiment could continue for an arbitrarily long time, which is impractical. Wald also studied the effect of setting an upper bound t_0 on the number of observations. If the sequential test does not have a decision by $t = t_0$, then we accept H_0 at the t_0-th trial when $\frac{p_{1,t_0}}{p_{0,t_0}} \leq 1$ and reject H otherwise. The truncation would cause inflation in errors and Wald advises that we should choose t_0 at least three times the expected number of observations required by the sequential test procedure to have negligible error inflation [20].

2.2 Composite hypotheses and sequential A/B tests

To study composite hypotheses with sequential analysis, Wald also came up with a mixture Sequential Probability Ratio test (mSPRT) for composite hypothesis $H_0 : \theta \in \theta_0$ against $H_1 : \theta \neq \theta_0$. It involves choosing a mixture distribution $w(\theta)$ for parameters in the region with preference of rejection and calculating the likelihood p_{1n} with

$$p_{1n} = \int_{\omega_r} f_\theta(x_{1:m})\omega(\theta)d\theta.$$

Often in A/B testing, we have two variants: control and treatment parametrized by θ_1 and θ_2 and the objective is to test whether treatment is different from control. Johari came up with always valid p-value for the power 1 ($\beta = 0$) mSPRT in [6] for $H_0 : \theta = \theta_1 - \theta_2 = 0$ vs $H_1 : \theta \neq 0$, and discussed the application to A/B tests for binary data using some normal approximations (see Section 4.3 in [5]).

Optional stopping in Bayesian A/B tests with the hypotheses above has been studied by Deng and Lu [3]. The Bayesian test controls False Discovery Rate and relies on using genuine priors with known prior odds.

2.3 Girshick Test for pairs of data

The mSPRT detects the existence of a treatment effect (that $\theta_1 \neq \theta_2$), but it does not make decisions about which group is better. This is in conflict with the primary goal of A/B testing: to choose the better variant. M.A. Girshick [4] proposed a special SPRT for choosing the better population. For single parameter models, suppose $X \sim f_{\theta_1}(\cdot)$ and $Y \sim f_{\theta_2}(\cdot)$, we want to test the hypothesis $H : \theta_1 \leq \theta_2$. This test assumes that data comes in pairs (x_i, y_i).

For a fixed choice of θ_1^0 and θ_2^0 we consider two hypotheses H_0: $\theta_1 = \theta_1^0, \theta_2 = \theta_2^0$ and $H_1 : \theta_1 = \theta_2^0, \theta_2 = \theta_1^0$ The values θ_1^0 and θ_2^0 define the magnitude of difference worth detecting for a business decision. The probability ratio test statistic with t pairs of data is

$$\frac{p_{1t}}{p_{0t}} = \prod_{i=1}^{t} \frac{f_{\theta_2^0}(x_i) f_{\theta_1^0}(y_i)}{f_{\theta_1^0}(x_i) f_{\theta_2^0}(y_i)}. \tag{3}$$

For Bernoulli models, this is the double dichotomy problem. If $X_i \overset{iid}{\sim} \text{Bern}(p_1)$ and $Y \overset{iid}{\sim} \text{Bern}(p_2)$, then a sufficient statistic at time t would be $(t\overline{X_t}, t\overline{Y_t})$, number of successes in each group. Then Equation 3 reduces to

$$\frac{p_{1t}}{p_{0t}} = \left(\frac{1 - p_2^0}{1 - p_1^0} \frac{p_1^0}{p_2^0} \right)^{t\overline{Y_t} - t\overline{X_t}}, \tag{4}$$

with log probability ratio

$$\log\left(\frac{p_{1t}}{p_{0t}}\right) = t\left(\overline{Y_t} - \overline{X_t}\right) \log\left(\frac{1 - p_2^0}{1 - p_1^0} \frac{p_1^0}{p_2^0} \right). \tag{5}$$

The log term we denote by

$$\nu(p_1, p_2) = \log\left(\frac{1 - p_2}{1 - p_1} \frac{p_1}{p_2} \right) \tag{6}$$

can be used as measure of deviance between p_1 and p_2. It satisfies the following properties: (1) $\nu(p_1, p_2) = 0$ if $p_1 = p_2$, (2) $\nu(p_1, p_2) < 0$ if $p_1 > p_2$ and (3) $\nu(p_2, p_1) = -\nu(p_1, p_2)$. This measure of deviance is visualized in Figure 1(a).

With any $\delta > 0$, this deviance measure can divide the parameter space $(0, 1)^2$ into three parts: $\omega_a = \{\nu(p_1, p_2) < -\delta\}$, $\omega_r = \{\nu(p_1, p_2) > \delta\}$ and $\omega_o = \{-\delta \leq \mu(p_1, p_2) \leq \delta\}$, corresponding to the region with preference for acceptance of H, region with preference for rejection of H and region of indifference in Wald's SPRT framework. We visualize these three regions defined by $\delta = 0.3$ in Figure 1(b).

Girshick's double dichotomy test goes as follows: fix some $\delta > 0$ and at time t, we would have t pairs of data and the log likelihood ratio is

$$Z_t = \log\left(\frac{p_{1t}}{p_{0t}}\right) = (-\delta) \times t \times \left(\overline{Y_t} - \overline{X_t}\right). \tag{7}$$

We can interpret the log probability ratio in Equation 7 as the product of the risk tolerance (δ), the sample size (t), and the difference in empirical averages $(\overline{Y_t} - \overline{X_t})$. This intuition becomes important for our imputed sequential Girshick test.

We calculate the log likelihood ratio then terminate and accept $H : p_1 < p_2$ at time t if $\log\left(\frac{p_{1t}}{p_{0t}}\right) \leq \log B$, terminate and reject H at time t if $\log\left(\frac{p_{1t}}{p_{0t}}\right) \geq \log A$, and continue the test by taking another observation pair if $\log B < \log\left(\frac{p_{1t}}{p_{0t}}\right) < \log A$. With A and B chosen based on α, β, this test has strength (α, β) because

$$\mathbb{P}(\text{reject } H | \omega_a) \leq \alpha \tag{8}$$

$$\mathbb{P}(\text{reject } H | \omega_r) \geq 1 - \beta \tag{9}$$

Again for all practical purposes we choose $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$.

It has been shown that the power of this test is constant for any p_1^0, p_2^0 such that $\nu(p_1^0, p_2^0) = d$. The exact power and distribution of stopping time $\tau = \inf\{t : Z_t \geq \log A \text{ or } Z_t \leq \log B\}$ are also given in [4].

2.4 Thompson sampling based controlled experiments

In addition to sequential analysis, multi-armed bandit experiments have also been used as an A/B testing procedure for best-arm identification in the online service economy [17]. The goal is to identify the best arm while simultaneously collecting the most reward in

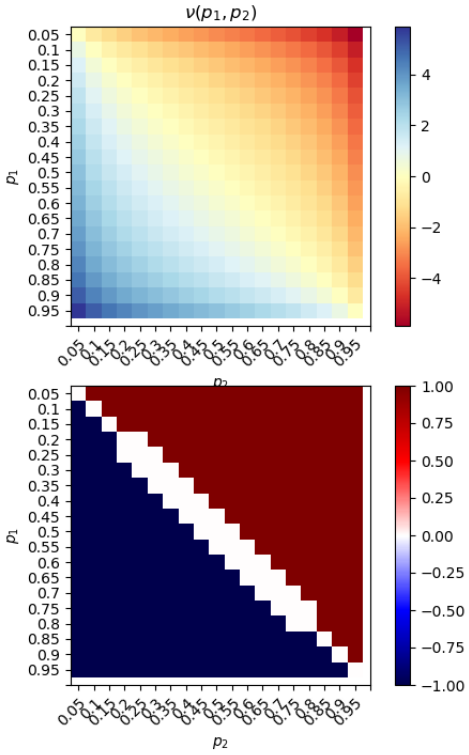

Figure 1: Top (a): the measure of deviance $\nu(p_1, p_2) = \log\left(\frac{1-p_2}{1-p_1}\frac{p_1}{p_2}\right)$ for binomial model. Bottom (b): The regions with different preferences of decisions for $H : p_1 < p_2$ with risk tolerance $\delta = 0.3$. The three regions are $\omega_a = \{\nu(p_1, p_2) < -\delta\}$ (the region with preference towards acceptance of H,in red) and $\omega_r = \{\nu(p_1, p_2) > \delta\}$ (region with preference towards rejection, in blue) and ω_i (region of indifference, in white). The Girshick test has Type-I error $\mathbb{P}(\text{reject } H|\omega_a) \leq \alpha$ and Type- II error $\mathbb{P}(\text{accept } H|\omega_r) \leq \beta$.

doing so. In this paper, we focus our discussion on Thompson sampling [18], as it is a default choice for many Internet companies, such as Google's Analytics Content Experiment platform [16]. Compared to fixed-time NHST, Thompson sampling based experiments are more cost-efficient because they gradually allocate users to the winning variant. These experiments are usually conducted in a streaming fashion and do not assume that observations in control and treatment groups come in pairs.

In it's most basic form, the Thompson sampling method works as follows: Suppose there are K groups that we want to compare, each having some average reward $\theta_1, \theta_2, \cdots, \theta_K$. Thompson sampling requires a prior on each θ_k. As data is collected, the posterior distribution of each θ_k is sequentially updated. After t data points $X_{1:t}$ are collected, the next customer is assigned to arm k based on the probability of the k-th arm being the optimal one, given the current data. This probability is calculated from the posterior distribution of rewards through

$$\mathbb{P}(\theta_k = \max\theta|X_{1:t}) = \int \mathbb{I}(\theta_k = \max\theta)\pi(\theta_{1:K}|X_{1:t})d\theta_{1:K}. \quad (10)$$

The performance of multi-armed bandit decision algorithms, such as Thompson sampling, is generally measured through regret. For a Bernoulli bandit problem, the per-period regret at t is $\text{regret}_t(\theta) = \max_k \theta_k - \theta_{a_t}$ where a_t indicates the allocation of the t-th sample. Suppose the process terminates at stopping time τ. Then the cumulative regret is

$$R(\tau) = \sum_{t=1}^{\tau}\left(\max_k \theta_k - \theta_{a_t}\right). \quad (11)$$

For more details of Thompson sampling, see [15, 17].

In terms of practicality, there have been theoretical studies on the asymptotic behaviors of Thompson sampling as well [8, 11]. Chapelle and Li [1] empirically evaluated the asymptotic regret of Thompson sampling and that it satisfies the Lai and Robins optimal bound [8].

How to properly stop a Thompson sampling based A/B test is still an open problem. On Google's platform, experiments are conducted for at least 2 weeks by default. Beyond that, the experiment is terminated when there is at least a 95% probability that the posterior expected value remaining in the experiment is less than 1% of the champion's conversion rate [16]. Johari also discussed using bandits within mSPRT based A/B tests [6] and it assumes the true value of $\frac{p_1+p_2}{2}$ is known.

While through Thompson sampling, we gain the ability to optimize reward by allocating more resources, users, etc. to the best variant, we lose the frequentist error control provided by sequential analysis. It is far more challenging to establish stopping conditions that allow for the control of Type-I&II error and when the data generating process varies over the course of the experiment. In the following, we will consider extensions to the Sequential Girshick Test with the goal of restoring some of this statistical control.

3 IMPUTED SEQUENTIAL GIRSHICK TEST

Traditional sequential tests, including the Girshick test, are not designed for streaming experiments. In a streaming environment, when a new customer visits the website at time t, she will be assigned into control or treatment group with some probability $\rho(t)$ and $1-\rho(t)$, resulting in unequal number of customers in each group. In this section, we propose modifications to the sequential Girshick test for pairs. This new test can support two allocation schemes: (1) static $\rho(t) = \rho$ and (2) data-dependent $\rho(t)$ with Thompson sampling.

3.1 Imputation in static allocation experiments

First, let us consider a static allocation rate experiment with $\rho(t) = \rho > 0$ for all t. Suppose at time t there have been m customers in the control group and $n = t - m$ customers in the treatment group. Instead of having t pairs of data, we only have t single observations. We can formulate this as a missing data problem with n responses missing from the control and m responses missing from the treatment.

As shown in Equation 4, it is sufficient to impute the total number of successes in the missing observations. We can impute the number of missing successes in control and in treatment with $n\hat{p_1} = \frac{n}{m}\sum_{i=1}^{m} x_i$ and $m\hat{p_2} = \frac{m}{n}\sum_{j=1}^{n} y_i$ respectively. Then the imputed log probability ratio for t pairs of data becomes

$$(-\delta)t\left(\frac{1}{n}\sum_{j=1}^{n} y_i - \frac{1}{m}\sum_{i=1}^{m} x_i\right) = (-\delta)t\left(\overline{Y_n} - \overline{X_m}\right). \tag{12}$$

If we treat t *imputed* pairs as t *observed* pairs, this imputation would inflate the Type-I& II errors above level (α, β), on top of the error inflation from setting $A = \frac{1-\beta}{\alpha}, B = \frac{\beta}{1-\alpha}$ and truncation at some t_0. So we also replace t in Equation 6 with an effective pair size $\frac{2}{\frac{1}{m}+\frac{1}{n}} = \frac{2mn}{t}$, the harmonic mean of m, n. With the imputations and effective pair size, the approximate log likelihood ratio becomes

$$\widehat{Z_t} = \widehat{\log\left(\frac{p_{1t}}{p_{0t}}\right)} = (-\delta)\frac{2mn}{t}\left(\overline{Y_n} - \overline{X_m}\right). \tag{13}$$

We would compare the approximate log likelihood ratio in Equation 13 against $\log A$ and $\log B$ to sequentially make decisions. Simulation experiments in Figure 2 demonstrate that this test indeed achieves strength (α, β).

By the law of large numbers, if not truncated with static allocation rate ρ, we have $\frac{mn}{t} \to \rho(1-\rho)t$ and $\overline{Y_n} - \overline{X_m} \to p_2 - p_1$ the true treatment effect as $t \to \infty$. If $p_1 \neq p_2$, the statistic $\widehat{Z_t} = O(t)$, and this process terminates in finite time with probability 1.

3.2 Thompson sampling and imputation

As mentioned in Section 2.4, there are no error-based stopping rules for Thompson sampling based hypothesis tests. In this section, we propose a stopping rule for such experiments and empirically evaluate the errors in Section 4.1

The likelihood ratio in Equation 13 does not explicitly involve the allocation ratio ρ, so why not use the same formulas with adaptive allocation methods, similar to Thompson sampling? As proved in the multi-armed bandit literature, Thompson sampling is greedy in the limit with infinite exploration [11], which means although every arms is visited infinitely often in the limit, eventually we allocate all the resources to the optimal arm. Assume control is the better variant in the Thompson sampling based A/B test, then $\rho(t) \to 1$ as $t \to \infty$.

Without truncation, there is a non zero probability that a process with Equation 13 never terminates, which we want to avoid with sequential tests. To this end, we use the geometric mean $\sqrt{mn}$ as the effective pair size for Thompson Sampling.

To approximate the treatment effect, we would still use the empirical average, although this estimator is consistent [11] but not unbiased [12]. The imputed log likelihood ratio test statistics for Thompson Sampling becomes

$$\widetilde{Z_t} = \widetilde{\log\left(\frac{p_{1,t}}{p_{0,t}}\right)} = (-\delta)\sqrt{mn}\left(\overline{Y_n} - \overline{X_m}\right). \tag{14}$$

We expect to see some minor error inflation with these imputations and they are visualized via simulations in Figure 3.

Designing a sequential test that accommodates multi-armed bandit based adaptive allocation is still an open problem as discussed by Johari in the discussions following Theorem 4 in [6]. It is challenging because the estimates of p_1 and p_2 are already biased with adaptive allocation [12]before introducing optional stopping. In the following section we will show via simulations that the test is able to reliably select the better variant while bounding errors as designed.

4 SIMULATED EXPERIMENTS

In this section we evaluate the imputed Girshick test procedure using three metrics: 1) reliability in terms of Type-I&II error control, 2) length of an experiment, and 3) cumulative regret during the experiment. All simulations in this section assume data from both the control and treatment groups to be drawn from Bernoulli models, Bern(p_1) and Bern(p_2), respectively.

4.1 Type I & II error

To test the significance level and power of the proposed imputed sequential Girshick test, we run experiments for various values of p_1, p_2 from the parameter space $(0, 1)^2$ with $\alpha = 0.05$ and $\beta = 0.05$ and truncation at $t_0 = 8000$.

Based on our simulation experiments with static $\rho = 0.5$ allocation, the imputed double dichotomy test using static allocation Equation 13 yields Type-I&II errors below α and β. Figure 2 shows the probability of accepting the hypothesis $H : p_1 < p_2$ in a heatmap. Each pixel is the average over 500 repetitions. The red lines are contour of the decision boundary $v(\theta_1, \theta_2) = \pm 0.3$. For all pairs of (p_1, p_2) tested such that $v(p_1, p_2) < -0.3$ (above the dotted red line) the probability of accepting H is greater than 0.95. For all values such that $v(p_1, p_2) > 0.3$ (below the solid red line), the probability of accepting H is below 0.05.

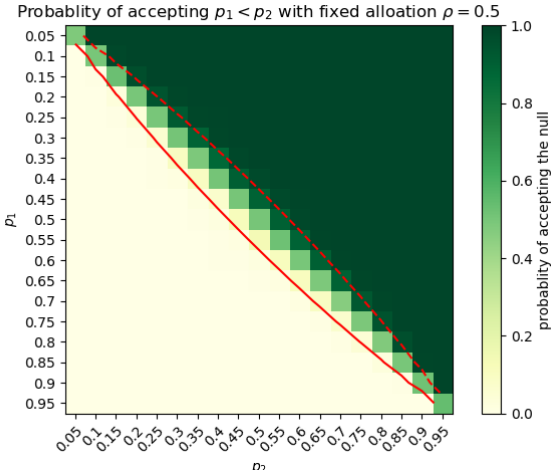

Figure 2: Probability of accepting the null hypothesis $H_0 : p_1 < p_2$ using the imputed Girshick test with fixed allocation ratio of $\rho = 0.5$ at significant level $\alpha = 0.05$, power $1 - \beta = 1 - 0.05$ from 100 repeated experiments. The red contour curves show the regions with preference of acceptance and rejection. They are chosen using $\frac{1-p_2}{1-p_1}\frac{p_1}{p_2} > 1/e^{\delta}$ or $< \frac{1}{e^{\delta}}$ with $\delta = 0.3$.

We also performed adaptive allocation experiments using the approximations in Equation 14 and independent Uniform priors on git a, and examined the probability of acceptance. As shown in Figure 3, there are some values of p_1, p_2 for which we observe higher Type-I and Type-II error rates than desired. These violation occur at values of p_1, p_2 in ω_a and ω_r that are very close to the decision boundary of $\nu(p_1, p_2) = \pm 0.3$. Typical values of these higher-than-expected errors rates are $0.06 - 0.08$ as compared to the expected 0.05. This suggests that in practice, if we want to have the flexibility of using adaptive allocations, we should set a conservative decision boundary δ close to 0.

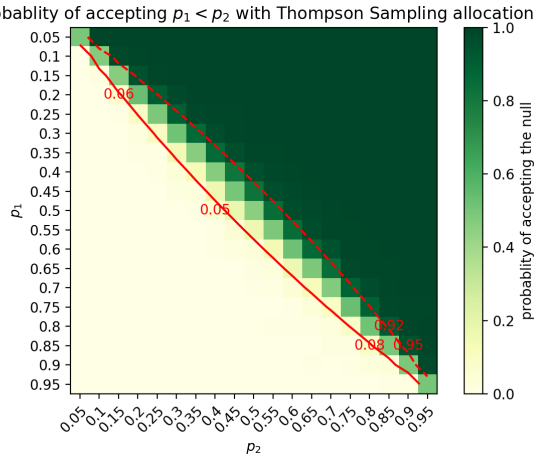

Figure 3: Probability of accepting the null hypothesis $H_0 : p_1 < p_2$ at significant level $\alpha = 0.05$, power $1 - \beta = 1 - 0.05$ of the imputed sequential Girshick test from 500 repeated experiments. The allocation scheme is Thompson sampling with the independent non-informative prior Beta$(1, 1)$ **on p_1 and p_2. The red contour curves shows the regions with preference of acceptance and rejection. They are chosen using $\frac{1-p_2}{1-p_1}\frac{p_1}{p_2} > e^{\delta}$ or $< \frac{1}{e^{\delta}}$ with $\delta = 0.3$. The values (p_1, p_2) leading to a error violation are marked with the acceptance probability in red.**

4.2 Stopping time

Another goal of sequential testing is to find statistically valid early stopping rules for A/B tests such that experimentation times can be shortened and product iterations can happen more rapidly. In the following simulation, we demonstrate the reliability of the procedure from the previous section (Figure 2 & 3) and study the stopping time of experiments to show that the total length of experiment time can be reduced.

For these simulations, we fix the data generating parameters at $p_1 = 0.45$ and $p_2 = 0.5$, and set the decision boundary at $\delta = 0.1$. We have $\nu(0.45, 0.5) = -0.20 < -\delta = -0.1$. This boundary is quite conservative, so we expect to see the desired error control with both fixed allocation and Thompson sampling allocation. The strength of the experiment is set to $\alpha = \beta = 0.05$ and truncation at $t_0 = 8000$. In this set up, we test different allocation schemes: (1) static allocation with $\rho = 0.5$, (2) static fixed allocation with $\rho = 0.7$, (3) Thompson sampling with independent Beta$(1, 1)$ prior on both p_1 and p_2 and (4) Thompson sampling with independent priors $p_1 \sim$ Beta$(45, 55)$ and $p_2 \sim$ Beta$(50, 50)$. The first set of priors are uniform, non-informative priors, which we would use without any prior knowledge. The second set of priors are informative priors that are consistent with the true values of the parameters. In practice,we cannot start with such good priors. Prior choice impacts the performance of Thompson sampling and in practice we cannot start with such ideal priors. W While prior specification is an important topic and is known to have a large impact on the finite-time performance of Thompson sampling, it is beyond the scope of the paper and interested readers can refer to Chapter 6.1 of [15].

	static allocation		Thompson sampling	
	$\rho = 0.5$	$\rho = 0.7$	Unif. priors	inform. priors
$\mathbb{P}(\text{accept}\|\omega_a)$	99.8 %	99.75%	97.7%	99.55%
average τ	1165.26	1383.86	1300.47	1537.59
min	186	148	263	235
median τ	1024	1194	1140	1376
max	5622	6214	4952	6329

Table 1: Comparison of number of observations required by the imputed Girshick test using different allocation schemes. For the same set up $p_1 = 0.45, p_2 = 0.5, \alpha = 0.05, \beta = 0.05$, a fixed-time two-sample proportion test needs 2589.479 observations in each group.

The results of the simulation are shown in Figure 4. As can be seen from the histogram, to achieve the same significance level of 0.05 and power of 0.95, the fixed-time proportion test requires 2589.479 pairs of samples (equivalent to 5178.958 total samples), whereas our sequential test only needs, on average, 1165 samples. In fact, all four tests rarely exceed the sample size of a fixed-time proportion test.

Similar to mixture Sequential Probability Ratio Tests (see Figure 2 in [6]) and Google's Analytic experiments (see Figure 3 in [16]), the distribution of number of observation required from a imputed sequential Girshick test is also right-skewed. Most of the time we arrive at the decision much earlier than a fixed-time NHST, but occasionally the experiment can take much longer. This is the price we pay for having the flexibility of a sequential test that supports both peeking and adaptive allocation.

4.3 Regret

Our third metric of quality for A/B tests is the total loss (of clicks, revenue, etc.) suffered over the course of an experiment, which we quantify as cumulative regret. For the same experiments mentioned in Table 1, we calculate the cumulative regret (Equation 11) at stopping time.

Using the imputed sequential Girshick test, we hope to see that Thompson sampling has a lower cumulative regret than static allocation, since introducing Thompson sampling is motivated by reducing cost. As our simulation experiments show, using Thompson sampling indeed gives lower cumulative regret compared to

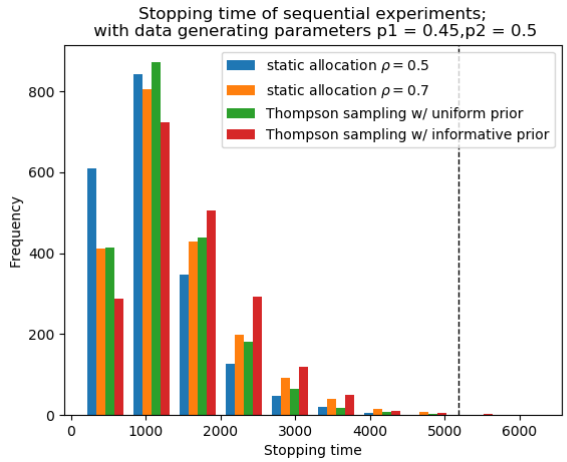

Figure 4: A histogram of stopping times for the imputed sequential Girshick test using different allocation schemes, corresponding to Table 1. The dashed black line is the sample size required by a fixed-time proportion test. There is a vanishingly small number of simulations where the sequential test requires more samples than the fixed-time proportion test.

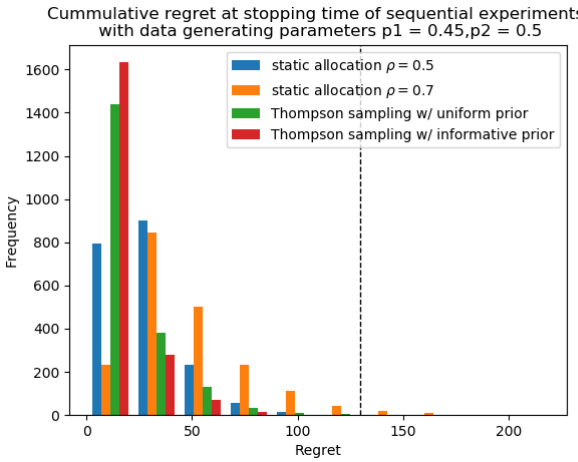

Figure 5: Histogram of cumulative regret of the 2000 simulation experiments in Table 1. Dashed black line is the cumulative regret of a fixed-time NHST.

	static allocation		Thompson sampling	
	$\rho = 0.5$	$\rho = 0.7$	Unif. priors	inform. priors
$\mathbb{P}(\text{accept}\|\omega_a)$	99.8%	99.75%	97.7%	99.55%
average R	29.15	48.46	17.56	13.64
min R	4.6	4.9	0.30	0.40
median R	25.75	41.85	11.38	9.00
max R	142.8	219.05	127.00	97.20

Table 2: Comparison of cumulative regret of imputed Girshick test using different allocation schemes. For the same set up $p_1 = 0.45, p_2 = 0.5, \alpha = 0.05, \beta = 0.05$, a fixed-time two-sample proportion test has cumulative regret 129.5.

static allocation experiment. Although Thompson sampling based experiments would run longer than static allocation, as seen in Figure 4, the cumulative regret is much smaller. In addition to that, all the sequential tests have smaller regret than the fixed-time test.

We also note that Thompson sampling with an informative prior (consistent with the truth) yields the best cumulative regret performance. Although in practice our prior specification may not be as good, the simulations indicate that Thompson sampling with an uninformative, uniform prior still has smaller regret than both static allocation experiments. This leads us to conclude that the superior regret performance will still hold even in light of a sub-optimal prior choice.

5 INDUSTRY EXPERIMENTS

To validate whether our theoretical considerations and insights from simulated experiments remain valid when applied to real experiments, we evaluate the performance of our sequential test on data from historical A/B tests on Etsy.com, a large e-commerce website specializing in handmade and vintage goods. Real-world experimental data introduces interesting challenges including non-stationary parameters and computational cost of real-time posterior updates.

5.1 Data and Experimental Set-up

We study a 12-day experiment that ran from 2018-05-15 to 2018-05-27 on Etsy's A/B testing platform. The metric of the experiment is hourly conversion rate per session. We model conversion rates in the control and treatment as p_1 and p_2 (respectively) with a Bernoulli model. The null hypothesis is $H : p_1 \leq p_2$, indicating that the treatment is better than the control, which is consistent with the outcome of this historical A/B test. Because many industry-scale experiments are tasked with detecting very small changes, we conservatively set $\delta = 0.01$. This will detect, for example, a conversion rate of 3.70% versus 3.75%.

We use bootstrap re-sampling to generate responses from the experiment and compare a static allocation and Thompson sampling allocation scheme, as we did with the simulated experiments conducted in Section 4. In all following experiments, we use an allocation rate $\rho = 0.5$. The Thompson sampling experiments used independent Uniform priors on p_1 and p_2. To update allocation probabilities in Equation 10, we use Monte Carlo approximations.

In contrast to the sequential updates where we calculate the log probability ratios and update posterior distributions at each step, we perform batch updates in these large-scale experiments. In the following experiments, we make a decision to terminate, continue, or to change the allocation ratio at every $b = 100$ steps. Using batch updates reduces computational cost in Thompson sampling and is easier to implement for streaming A/B tests with frequent user visits.

5.2 Results

In this section, we evaluate these bootstrapped experiments using the same metrics discussed in Section 4, namely error control, length of experiment, and cumulative regret.

In terms of error control, 100% of static allocation experiments and 98% of Thompson sampling experiments correctly accepted the hypothesis $H : p_1 \leq p_2$. When evaluating by length of experiment

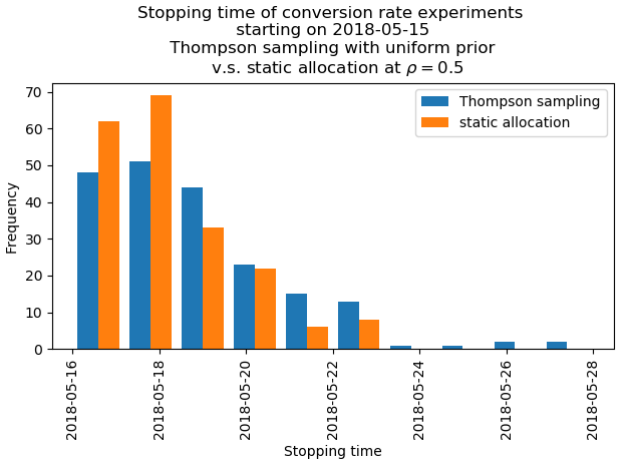

Figure 6: Histogram of stopping times of 200 bootstrap re-sampling experiments. Using static sampling, 69/200 experiments stop on 2018-05-18 and the longest experiment would end on 2018-05-18. Thompson sampling experiment take longer, 51/200 experiments end on 2018-05-18 with the longest experiment stopping on 2018-05-28.

time, most experiments using both allocation schemes would stop on 2018-05-18 (the 4th day). This is in contrast to the original end date of 2018-05-27 of the historical experiment. A histogram of the stopping time of this conversion rate experiment is shown in Figure 6. We note that some Thompson sampling based experiments can last longer, with 1/200 bootstrapped experiments ending after the original end date of 2018-05-27.

Cumulative regret, the third metric for evaluation, from both allocation schemes are shown in Figure 7. The average regret from Thompson sampling and static allocation are 398 and 288, respectively.

This outcome differs from our simulated experiments, specifically in which static allocation at $\rho = 0.5$ outperforms Thompson sampling with uniform priors. However, recall that cumulative regret is the average number of 'successes' we lose in an experiment. Considering the scale of the experiments, this is very low cost. Furthermore, to explain this discrepancy, we speculate that the batch updates contribute to the higher regret from Thompson sampling. Updating at every $b = 100$ samples means holding onto newly collected evidence. But the sooner we use the information, the higher the conversion and revenue the system will have. Also, the use of a Uniform prior could cause Thompson sampling to under-perform static allocation. For a conversion rate experiment, Uniform prior is a missspecified prior since it is not coherent with

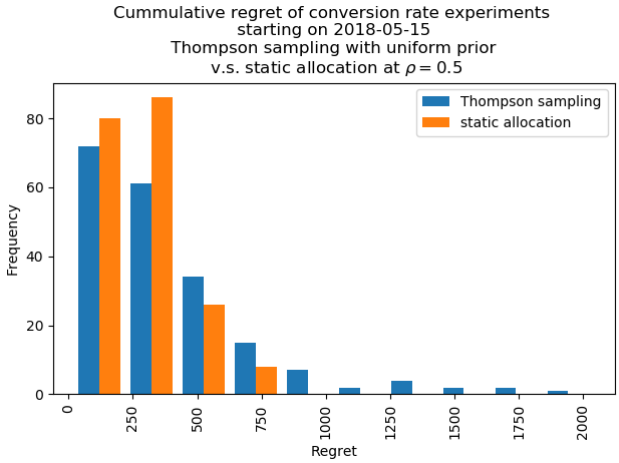

Figure 7: Histogram of cumulative regret of the 200 bootstrap re-sampling experiments. The average regret from Thompson sampling is 398 and average regret 288 from static allocation.

typical range of conversion rates. That missspecified priors delay learning is illustrated in our simulation experiments and has been noted in literature [15]. Practically speaking, we could incorporate informative priors given access to more historical experiments, because they capture the typical range of conversion rates. Precedents of learning objective priors from historical experiments have been seen in [2]. This may close the gap in terms of regret performance between static allocation and Thompson sampling experiments.

6 DISCUSSION

Time and opportunity cost during A/B tests are two major challenges of Internet experiments. As we have demonstrated with simulations (Section 4) and real-data experiments (Section 5), static allocation experiments ends early while Thompson sampling based allocation would save cumulative experimental cost if we have good priors.

6.1 Practical considerations

The proposed imputed sequential Girshick test procedure allows enough flexibility like static/adaptive allocation, peeking and batch updates. Like all sequential tests, the price we have to pay is occasionally waiting longer than a fixed-time `NHST` with the same strength.

With the benefits of sequential tests presented in this paper, we want to acknowledge that continuously monitoring is not always recommended. The current framework does not support time-varying trends in the data or delays in response for example in monthly or seasonal metrics.

6.2 Future directions

The superior regret performance of Thompson sampling applies in the asymptotic setting but with sequential tests we would not enter the asymptotic regime. There is very little theoretical results

on finite time regret [7] of Thompson sampling at some proper stopping time.

Although fixed-time experiments are expensive, it does offer the benefit of unbiased estimation of parameters. The empirical averages from imputed sequential Girshick test are biased estimates of the true parameters, because of the decisions we make to stop the test early. Post-selection inference has recently been discussed by many authors, for example for forward step-wise regression [19] and Lasso [9] and for additive total effects in A/B tests by Lee and Shen from Airbnb [10].

There is also negative bias in treatment effect estimation from using Thompson sampling. Estimation bias from adaptively collected data have been studied by Nie et al. [12], with debiasing for Thompson sampling in the Gaussian setting. Bias correction from Thompson sampling for Beta-Bernoulli is still an open problem. We expect that using bias adjusted estimates of p_1 and p_2 in Equation 14 should have better performance in terms of error control and costs.

In the future we would like to extend this test procedure to other models for example normal distribution with known variance and regret in terms of profit during experiments. We would also like to study an extension of the Girshick test for A/B/n experiments.

REFERENCES

[1] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.

[2] Alex Deng. 2015. Objective bayesian two sample hypothesis testing for online controlled experiments. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 923–928.

[3] Alex Deng, Jiannan Lu, and Shouyuan Chen. 2016. Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. IEEE, 243–252.

[4] Meyer Abraham Girshick. 1946. Contributions to the theory of sequential analysis. I. *The Annals of Mathematical Statistics* (1946), 123–143.

[5] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at A/B Tests: Why It Matters, and What to Do About It. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 1517–1525. https://doi.org/10.1145/3097983.3097992

[6] Ramesh Johari, Leo Pekelis, and David J Walsh. 2015. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922* (2015).

[7] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*. Springer, 199–213.

[8] Tze Leung Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.

[9] Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. 2016. Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44, 3 (2016), 907–927.

[10] Minyong R Lee and Milan Shen. 2018. Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 491–499.

[11] Benedict C May, Nathan Korda, Anthony Lee, and David S Leslie. 2012. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* 13, Jun (2012), 2069–2106.

[12] Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. 2018. Why Adaptively Collected Data Have Negative Bias and How to Correct for It. In *International Conference on Artificial Intelligence and Statistics*. 1261–1269.

[13] Herbert Robbins. 1970. Statistical Methods Related to the Law of the Iterated Logarithm. *The Annals of Mathematical Statistics* 41, 5 (10 1970), 1397–1409. https://doi.org/10.1214/aoms/1177696786

[14] H. Robbins and D. Siegmund. 1974. The Expected Sample Size of Some Tests of Power One. *The Annals of Statistics* 2, 3 (1974), 415–436. http://www.jstor.org/stable/2958130

[15] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning* 11, 1 (2018), 1–96.

[16] Steven L Scott. 2012. Overview of Content Experiments: Multi-armed bandit experiments. https://support.google.com/analytics/answer/2844870?hl=en&ref_topic=1745207

[17] Steven L Scott. 2015. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry* 31, 1 (2015), 37–45.

[18] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.

[19] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. 2016. Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* 111, 514 (2016), 600–620.

[20] Abraham Wald. 1945. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics* 16, 2 (1945), 117–186.