Peeking at A/B Tests

Why it matters, and what to do about it

Ramesh Johari*
Stanford University
rjohari@stanford.edu

Pete Koomen†
Optimizely, Inc.

Leonid Pekelis‡
Optimizely, Inc.
lpekelis@gmail.com

David Walsh§
Stanford University
dwalsh@stanford.edu

ABSTRACT

This paper reports on novel statistical methodology, which has been deployed by the commercial A/B testing platform Optimizely to communicate experimental results to their customers. Our methodology addresses the issue that traditional p-values and confidence intervals give unreliable inference. This is because users of A/B testing software are known to *continuously monitor* these measures as the experiment is running. We provide *always valid* p-values and confidence intervals that are provably robust to this effect. Not only does this make it safe for a user to continuously monitor, but it empowers her to detect true effects more efficiently. This paper provides simulations and numerical studies on Optimizely's data, demonstrating an improvement in detection performance over traditional methods.

KEYWORDS

A/B testing, sequential hypothesis testing, p-values, confidence intervals

1 INTRODUCTION

Web applications typically optimize their product offerings using randomized controlled trials (RCTs); in industry parlance this is known as *A/B testing*. The rapid rise of A/B testing has led to the emergence of a number of widely used platforms that handle the implementation of these experiments [10, 20]. The typical A/B test compares the values of a parameter across two variations (*control* and *treatment*) to see if one variation offers an opportunity to improve their service, while the A/B testing platform communicates results to the user via standard frequentist parameter testing measures, i.e., p-values and confidence intervals. In doing so, they obtain a very simple "user interface", because these measures isolate the task of analyzing experiments from the details of their design and implementation.

Crucially, the inferential validity of these p-values and confidence intervals requires the separation between the design and analysis of experiments to be strictly maintained. In particular, the sample size must be fixed in advance. Compare this to A/B testing practice, where users often *continuously monitor* the p-values and confidence intervals reported in order to re-adjust the sample size of an experiment dynamically [14]. Figure 1 shows a typical A/B testing dashboard that enables such behavior.

This "peeking" behavior results because the opportunity cost of longer experiments is large, so there is value to detecting true effects as quickly as possible, or giving up if it appears that no effect will be detected soon so that the user may test something else. Further, most users lack good prior understanding of both their tolerance for longer experiments as well as the effect size they seek, frustrating attempts to optimize the sample size in advance. Peeking early at results to trade off maximum detection with minimum samples dynamically seems like a substantial benefit of the real-time data that modern A/B testing environments can provide.

Unfortunately, stopping experiments in an adaptive manner through continuous monitoring of the dashboard will severely *favorably bias* the selection of experiments deemed significant. Indeed, very high false positive probabilities can be obtained—well in excess of the nominal desired false positive probability (typically set at 5%). As an example, even with 10,000 samples (quite common in online A/B testing), we find that the false positive probability can easily be inflated by 5-10x. That means that, throughout the industry, users have been drawing inferences that are not supported by their data.

Our paper presents the approach taken to address this challenge within the large-scale commercial A/B testing platform Optimizely. We develop novel methodology to compute p-values and confidence intervals; our measures, which we call *always valid*, allow users to continuously monitor the experiment and stop at a data-dependent time of their choosing, while maintaining control over false positive probability at a desired pre-set level. This protects statistically naive users, and lets all users leverage real-time data to trade off the detection power and sample size dynamically. As described in the paper, our methods build on classical results in the sequential testing literature in statistics. The methods we describe were implemented in the Optimizely platform in January 2015 as *Optimizely Stats Engine*, and have been in use across all products including mobile,

*RJ is a technical advisor to Optimizely, Inc.; this work was completed as part of his work with Optimizely.
†PK is co-founder and Chief Technology Officer of Optimizely, Inc.
‡LP is a technical advisor to Optimizely, Inc.; this work was completed when he was an employee at Optimizely.
§This work was completed while DW was employed by Optimizely, Inc.

KDD'17, August 13–17, 2017, Halifax, NS, Canada.
 ACM ISBN 978-1-4503-4887-4/17/08.
DOI: http://dx.doi.org/10.1145/3097983.3097992

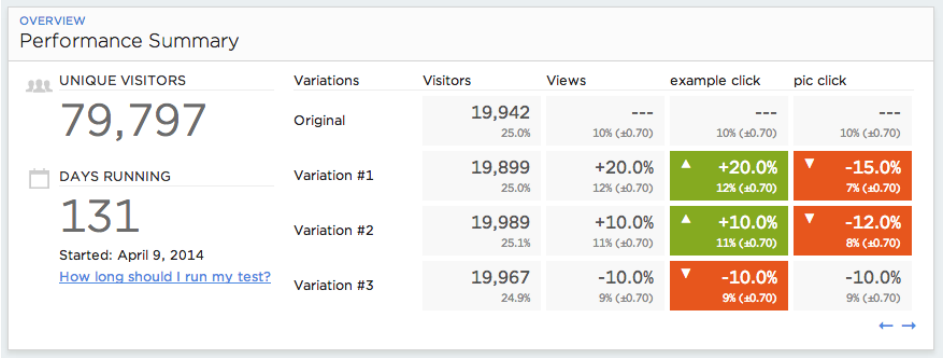

Figure 1: A typical results page in Optimizely's A/B testing dashboard. The dashboard encourages users to *continuously monitor* their experiments, by providing updated results on experiments in real-time.

web, and server-side testing; hundreds of thousands of experiments have been run by thousands of customers since its launch.

In Section 2, we outline the basic A/B testing problem, as well as the typical approach used today. In Section 3, we discuss why continuous monitoring "breaks" the existing paradigm and leads to invalid inference, and we propose a definition of always valid p-values and confidence intervals that admit valid inference despite continuous monitoring of tests by users. In Section 4, we give the approach taken in the Optimizely platform to compute these measures; in particular, they are derived from a novel generalization of the *mixture sequential probability ratio test* (mSPRT) [16]. In Section 5, we empirically demonstrate that our approach both allows users to control false positive probability, *and* improves the user's ability to trade off between detection of real effects and the length of the experiment (in an appropriate sense).

The core of our solution is formulated for the basic A/B testing problem with two variations (treatment and control). We conclude in Section 6 by addressing challenges that arise for *multivariate testing*, where users have many variations and metrics of interest that they compare simultaneously. Multivariate testing immediately gives rise to a severe *multiple comparisons* problem, where users can overinterpret significant results unless appropriate corrections are applied [21]. In our deployment, always valid p-values are combined with multiple hypothesis testing correction procedures to provide a robust inference platform for experimenters, supporting both continuous monitoring and multivariate testing.

2 PRELIMINARIES

In this section, we describe the typical approach for analyzing A/B tests based on the frequentist theory of hypothesis testing, which we refer to as *fixed-horizon testing*.

2.1 Experiments and decision rules

We begin by introducing two benchmark settings that we employ throughout the paper: experiments that involve testing one variation against a known baseline, and experiments that compare two variations against each other. The former is used to motivate our technical approach; the latter is the scenario encountered in A/B testing.

One-variation experiment. In a *one-variation experiment*, we test a single variation (or treatment) against a known baseline. In particular, we suppose independent observations from an exponential family $\mathbf{X} = (X_n)_{n=1}^{\infty} \overset{iid}{\sim} F_\theta$, where the parameter θ takes values in $\Theta \subset \mathbb{R}^p$. In this setting, we consider the problem of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta \neq \theta_0$. Here θ_0 is the known baseline of comparison.

Throughout the paper, we index probability distributions by the parameters; e.g., $\mathbb{P}_\theta$ denotes the probability distribution on the data induced by parameter θ.

Two-variation experiment. In a *two-variation experiment* or A/B test, we test two variations (e.g., treatment and control, or A and B) against each other. Here we observe two independent i.i.d. sequences $\mathbf{X}$ and $\mathbf{Y}$, corresponding to the observations on visitors receiving experiences A and B respectively. In studying A/B tests, we restrict the data model to the two most common cases encountered in practice: Bernoulli data with success probabilities μ^A and μ^B (used to model binary outcomes such as clicks, conversions, etc.); and normal data with means μ^A and μ^B and known variance σ^2 (used to model continuous-valued outcomes such as time on site). In this setting, we consider the problem of testing the null hypothesis $H_0 : \theta := \mu^B - \mu^A = 0$ against $H_1 : \theta \neq 0$.

Decision rules. The experimenter needs to decide how long to run the test, and whether to reject the null hypothesis when the test is done. We formalize this process through the notion of a *decision rule*. Formally, a decision rule is a pair (T, δ), where T is a *stopping time* that denotes the sample size at which the test is ended, and δ is a binary-valued decision dependent only on the observations up to time T, where $\delta = 1$ indicates that H_0 is rejected.[1] A stopping time is any time that is dependent only on the data observed up to that time; therefore, this definition captures the crucial feature of decision-making in A/B tests that *the terminal sample size may be data-dependent*.

Note that we allow the possibility that $T = \infty$. This formalism is allowed to capture the idea that in advance, we do not know how long a user would be willing to run a test. Of course in practice,

[1]Even more formally, let $(\mathcal{F}_n)_{n=1}^{\infty}$ denote the filtration generated by observations up to time n. Then in a decision rule, T must be a stopping time with respect to $(\mathcal{F}_n)_{n=1}^{\infty}$, and δ must be a $(\mathcal{F}_T)$-measurable binary random variable.

users will not run tests indefinitely. Accounting for this fact in our inferential process plays an important role below. If $T = \infty$, we assume $\delta = 0$: if the test runs forever then *de facto*, the null is never rejected.

2.2 Fixed-horizon testing

In this subsection we recap the approach typically used to run and analyze A/B tests today, based on frequentist hypothesis testing.

The textbook approach for experimental design, dating back to the seminal work of R.A. Fisher [5], is as follows:

Step 1: Commit to a fixed sample size n. This is a crucial point: the statistical measures typically used to analyzed A/B tests are computed under the presumption that the sample size was fixed *in advance*. We refer to this approach as *fixed-horizon* testing.

Step 2: Choose a desired false positive probability α. Next, the user chooses a desired control on the probability of *Type I error* or a false positive, i.e., the probability under the null hypothesis of an erroneous rejection. It is typical to use $\alpha = 0.05$ in practice, i.e., a desired significance level of $1 - \alpha = 95\%$.

Step 3: Collect n observations and compute the appropriate test statistic. Preferred test statistics are ones which can control Type I error with high *power* at each possible alternative: the probability of correctly rejecting the null hypothesis, after having received n observations.

Indeed, a well known result is that, in a one-variation experiment, for data in an exponential family, there exist a family of uniformly most powerful (UMP) test statistics, τ_n, with decision rules: reject the null hypothesis if τ_n exceeds a threshold $k(\alpha)$ (see, e.g., [13], Chapter 4). Perhaps the most common examples are the one- and two-sample z-tests and t-tests, used for data assumed to arise from a normal distribution with known or unknown variance respectively.

Step 4: Compute a p-value p_n, and reject the null hypothesis if $p_n \leq \alpha$. Informally, the *p-value* is *the probability under the null hypothesis of finding data at least as extreme as the observed test statistic.* If the p-value is small, that is considered as evidence that the null hypothesis is likely false.

Formally, we define the p-value at n as the smallest α such that the α-level UMP decision rule rejects the null hypothesis:

$$p_n = \inf\{\alpha : \tau_n \geq k(\alpha)\}.$$

Since the p-value was computed assuming a fixed sample size n, we refer to this as a *fixed-horizon p-value.*

Observe that the rule to reject when $\tau_n \geq k(\alpha)$ controls false positive probability at level α. But this is only possible if, under the null hypothesis, the event $\tau_n \geq k(\alpha)$ occurs with probability no greater than α. This is the sense in which the p-value captures the probability of finding data as extreme as the test statistic under the null hypothesis.

The last point (Step 4) is in large measure a reason for the popularity of this paradigm, despite the obvious subtleties in appropriate interpretation of frequentist hypothesis tests. The decision-making process using p-values is remarkably *simple*: it does not require the user to understand any of the intricacies of the procedure that led to the test statistic, and instead summarizes the outcome of the experiment in a number that can be directly compared to the desired probability of false positives. Further, p-values have a natural *transparency* property: different individuals can have higher or lower levels of α (corresponding to being less or more conservative), and can make statistically valid decisions using the same observed p-value.

We conclude with a definition that we employ in our later technical development. That the family of UMP decision rules controls Type I error may be stated as the following validity property on p_n for a one-variation experiment:

$$\forall s \in [0,1],\ \mathbb{P}_{\theta_0}(p_n \leq s) \leq s; \tag{1}$$

i.e., at fixed n, under the null hypothesis, the p-value is *superuniform*. For a two-variation experiment, the same validity condition is the following:

$$\forall \mu^A = \mu^B, s \in [0,1],\quad \mathbb{P}_{\mu^A,\mu^B}(p_n \leq s) \leq s. \tag{2}$$

We refer to a sequence of p-values that satisfy (1) or (2) as a *fixed-horizon p-value process.*

2.3 Confidence intervals

Using a standard duality between p-values and confidence intervals, the same approach can be used to construct confidence intervals as well. In particular, consider the family of fixed-horizon tests $\delta_n(\alpha)$ for testing $H_0 : \theta = \tilde{\theta}$ for each $\tilde{\theta} \in \Theta$. The $1 - \alpha$ confidence interval I_n is the set of $\tilde{\theta}$ that are *not* rejected. If the fixed-horizon test controls Type I error, that translates into the following coverage bound on the confidence interval:

$$\forall \theta \in \Theta,\ \mathbb{P}_\theta(\theta \in I_n) \geq 1 - \alpha. \tag{3}$$

More generally, we call any data-dependent interval I_n that satisfies the preceding bound a $1 - \alpha$ confidence interval for θ. (The same definition generalizes to two-variation experiments as well.)

Because of this duality, our technical development primarily focuses on p-values; the corresponding hypothesis tests can be used to construct confidence intervals in the preceding manner.

3 ALWAYS VALID INFERENCE

Unfortunately, a key failure mode of the experimental approach described in Section 2.2 is that it requires the user to commit to the sample size *in advance* of running the experiment. This is the property that allows the use of optimal UMP decision rules in the first place.

On the other hand, while they maximize power for the given n, the power increases as n is increased, and so the user must choose n to trade off power against the opportunity cost of waiting for more samples. A key feature of modern A/B testing platforms is precisely that they enable the user to *continuously monitor* experiments, allowing the user to adaptively adjust this trade off based on the observed data. This behavior leads to favorable biasing of the sample paths on which the user rejects the null hypothesis, which in turn leads to substantial inflation of the false positive probability.

Figure 2 illustrates the issue. There we simulate data in an A/B test where in fact both treatment and control consist of Normal(0, 1) data; we then test the null hypothesis that the mean is the same in both variations. The three curves show the realized Type I error if the null hypothesis is rejected the *first time* the p-value falls below

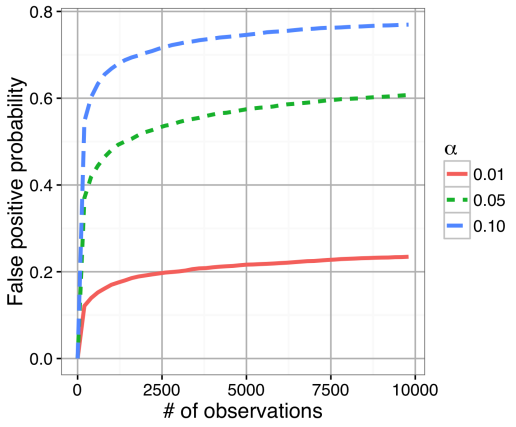

Figure 2: Type I error vs. run length with continuous monitoring. The curves show the realized false positive probability when the user rejects the first time the p-value crosses the given α level.

the given level of α; as can be seen, this data-dependent decision rule using fixed-horizon p-values massively inflates Type I error.[2]

It can be shown theoretically that *any* fixed level of α is *guaranteed* to be crossed by the p-value under the null hypothesis, if the experimenter waits long enough [19]. In other words, if the null hypothesis is rejected the first time the p-value crosses α, with increasing data the false positive probability approaches 100%!

This section outlines a statistical approach that corrects for this problem. Fundamentally, our premise is that users are right to want to adaptively determine their sample size: there are opportunity costs to longer experiments. We ask: how can we allow a user to stop when they wish, while still controlling the false positive probability at the level α?

3.1 Always valid p-values

The first significant contribution of our paper is the definition of *always valid* p-values that control Type I error, no matter when the user chooses to stop the test. These protect against adaptive data-dependent choices of the sample size, and let the user trade off detection power and sample size dynamically as they see fit. We define always valid p-values for one-variation experiments; the definitions extend naturally to the uniform validity required for two variations.

Definition 3.1. A sequence of fixed-horizon p-values (p_n) is an *always valid p-value process* if given any (possibly infinite) stopping time T, there holds:

$$\forall s \in [0,1],\ \mathbb{P}_{\theta_0}(p_T \le s) \le s. \tag{4}$$

The key difference from the condition (1) is the following: now we imagine the user observes data as it arrives, and makes a choice of when to stop the experiment; this is the data-dependent stopping time T. The user *then* observes the p-value at this time. Our requirement is that even though the p-value is viewed at a time that is data-dependent, Type I error must still be controlled.

[2]The same occurs even if slightly more sophisticated approaches are used. For example, suppose that the test is stopped the first time the p-value falls below $\alpha = 0.05$, and the calculated power at the observed effect size is above a fixed threshold (in this case, 0.8). This is known as a "post-hoc" power calculation [7]. This approach turns out to be equivalent to rejecting the first time the p-value falls below a given, lower value of α, and also significantly inflates the false positive probability.

3.2 Sequential tests and always validity

In statistics, decision rules where the terminal sample size is allowed to be data-dependent are commonly referred to as *sequential tests*. In this section, we show how sequential tests can be used to construct always valid p-values, and vice versa. Sequential analysis is a mature field with roots dating back to [22], and it has gained recent popularity for online experimentation [15] [1], especially in connection to multi-armed bandits [18]. For history, methodology, and theory, we direct the reader to the encyclopedic resource of [6].

Formally, a sequential test is a nested family of decision rules $(T(\alpha), \delta(\alpha))$ parameterized by their Type I error rates $0 < \alpha < 1$ with the following properties:

(1) Each decision rule controls Type I error at the stated level : $\mathbb{P}_{\theta_0}(\delta(\alpha) = 1) \le \alpha$.
(2) The decision rules are *nested*: $T(\alpha)$ is (almost surely) nonincreasing in α, and $\delta(\alpha)$ is (almost surely) nondecreasing in α. That is, the less conservative rules necessarily terminate faster and make more rejections.

The following theorem shows that sequential tests and always valid p-values are closely related; the proof can be found in the Appendix.

Theorem 3.2.

(1) *Let $(T(\alpha), \delta(\alpha))$ be a sequential test. Then*

$$p_n = \inf\{\alpha : T(\alpha) \le n, \delta(\alpha) = 1\} \tag{5}$$

defines an always valid p-value process.

(2) *For any always valid p-value process $(p_n)_{n=1}^{\infty}$, a sequential test $(\tilde{T}(\alpha), \tilde{\delta}(\alpha))$ is obtained from $(p_n)_{n=1}^{\infty}$ as follows:*

$$\tilde{T}(\alpha) = \inf\{n : p_n \le \alpha\}; \tag{6}$$

$$\tilde{\delta}(\alpha) = \mathbf{1}\{\tilde{T}(\alpha) < \infty\}. \tag{7}$$

(3) *Let $(T(\alpha), \delta(\alpha))$ be any sequential test where $T = \infty$ whenever $\delta = 0$. If $(p_n)_{n=1}^{\infty}$ is derived as in (5), then the construction (6)-(7) recovers the original sequential test: $(\tilde{T}(\alpha), \tilde{\delta}(\alpha)) = (T(\alpha), \delta(\alpha))$.*

Note that part (3) says the simple rule "reject when the p-value is $\le \alpha$" implements the original sequential test. While the p-value defined in part (1) of the theorem is not unique for satisfying part (3), it is the unique such process that is (almost surely) *monotonically nonincreasing* in n, which increases interpretability (see Section 4.4).

3.3 Confidence intervals

We conclude by extending always validity to confidence intervals. Always valid confidence intervals may be constructed from always valid p-values just as in the fixed-horizon context.

Definition 3.3. A sequence of fixed-horizon $(1-\alpha)$-level confidence intervals (I_n) is an *always valid confidence interval process* if

for any stopping time T, the corresponding interval I_T has $1-\alpha$ coverage of the true parameter: for all $\theta \in \Theta$, $\mathsf{P}_\theta(\theta \in I_T) \geq 1-\alpha$.

The following proposition follows immediately from the definitions.

Proposition 3.4. *Suppose that, for each $\tilde{\theta} \in \Theta$, $(p_n^{\tilde{\theta}})$ is an always valid p-value process for the test of $\theta = \tilde{\theta}$. Then $I_n = \left\{\theta : p_n^\theta > \alpha\right\}$ is an always valid $(1-\alpha)$-level CI process.*

4 CONSTRUCTING ALWAYS VALID P-VALUES

In light of Theorem 3.2, one might ask: if always valid p-values can be constructed from any sequential test, what value do these p-values offer over implementing the best sequential test for the experimenter directly? In particular, given the user's choice of Type I error constraint α and her desired balance between power and sample size (suitably defined, cf. [6]), we could choose a sequential test from the literature that optimizes her objective. But for an A/B testing platform, there is a fundamental problem with this approach: *each user wants a different trade off between power and sample size!* Some users are willing to wait longer than others. Always valid p-values let the user make the trade off for herself.

In this section, we describe the *particular* family of sequential tests we use to construct always valid p-values in the Optimizely platform: the *mixture sequential probability ratio test* (mSPRT). In Section 5, we show that mSPRTs are a judicious choice that enables users with heterogeneous needs to trade off power and sample size effectively.

4.1 The mixture sequential probability ratio test (mSPRT)

mSPRTs have been studied in the literature for one-variation experiments [16]; our extension to two-variation experiments is described below. The test is defined by a "mixing" distribution H over Θ, where H is assumed to have a density h that is positive everywhere. Using H, we first compute the following mixture of likelihood ratios against the null hypothesis that $\theta = \theta_0$:

$$\Lambda_n^{H,\theta_0} = \int_\Theta \prod_{m=1}^{n} \frac{f_\theta(X_m)}{f_{\theta_0}(X_m)} h(\theta) d\theta. \tag{8}$$

Intuitively, Λ_n^{H,θ_0} represents the evidence against H_0 in favor of a mixture of alternative hypotheses, based on the first n observations.

Now the mSPRT is fairly simple: given a desired false positive probability α, it stops and rejects the null hypothesis at the first time $T = T^H(\alpha)$ that $\Lambda_T^{H,\theta_0} \geq \alpha^{-1}$; if no such time exists, it never rejects the null hypothesis. Using standard martingale techniques, it can be shown that this sequential test controls Type I error at level α [19].[3]

[3] The basic idea is to observe that under the null hypothesis, the likelihood ratio at any θ is a martingale, and therefore Λ_n^{H,θ_0} is also a martingale. The result then follows by applying the optional stopping theorem.

4.2 p-values and confidence intervals from the mSPRT

We convert the mSPRT to always valid p-values and confidence intervals using Theorem 3.2 and Proposition 3.4. In particular, suppose we are given the sequence Λ_n^{H,θ_0}. Then by Theorem 3.2, the associated always valid p-values are seen to satisfy the following simple recursion:

$$p_0 = 1; \quad p_n = \min\{p_{n-1}, 1/\Lambda_n^{H,\theta_0}\}. \tag{9}$$

In particular, note that this means always valid p-values can be easily computed in a *streaming* fashion, making them amenable to implementation in a real-time A/B testing dashboard.

Applying Proposition 3.4 to these p-values, we find that always valid confidence intervals are given by the following recursion:

$$I_0 = \Theta, I_n = I_{n-1} \cap \{\tilde{\theta} : \Lambda^{(H,\tilde{\theta})} \geq \alpha^{-1}\}. \tag{10}$$

For data generated by general exponential families, as long as an appropriate conjugate prior is chosen as H, computation of Λ_n^{H,θ_0} (and thus both always valid p-values and always valid confidence intervals) is inexpensive. For data generated from a normal distribution (i.e., where $F_\theta = N(\theta, \sigma^2)$) it turns out that if we use a normal mixing distribution centered at the null hypothesis (i.e., $H = N(\theta_0, \tau^2)$, then we obtain a closed form formula for Λ_n^{H,θ_0}:

$$\Lambda_n^{H,\theta_0} = \frac{\sigma}{\sqrt{\sigma^2 + n\tau^2}} \exp\left\{\frac{n^2\tau^2(\bar{X}_n - \theta_0)^2}{2\sigma^2(\sigma^2 + n\tau^2)}\right\}.$$

(Here $\bar{X}_n$ is the sample mean up to n.) This formula can then be used to compute both always valid p-values and confidence intervals in a streaming format.

4.3 The mSPRT for A/B tests

The second major contribution of our paper is a novel generalization of the mSPRT to A/B tests. To get an mSPRT for A/B testing, we need to define a mixture likelihood ratio $\tilde{\Lambda}_n^{H,\theta_0}$ for two-variation experiments, as a function of the data $X_1, \ldots, X_n, Y_1, \ldots Y_n$.

We start by considering normal data. In this case, note that for any μ^A and μ^B, $Z_n = Y_n - X_n \sim N(\theta, 2\sigma^2)$. We can thus simply apply the one-variation mSPRT to the sequence $\{Z_n\}$; this leads to the following definition:

$$\tilde{\Lambda}_n^{H,\theta_0} = \sqrt{\frac{2\sigma^2}{2\sigma^2 + n\tau^2}} \exp\left\{\frac{n^2\tau^2(\bar{Y}_n - \bar{X}_n - \theta_0)^2}{4\sigma^2(2\sigma^2 + n\tau^2)}\right\}, \tag{11}$$

where θ_0 is the difference of means in the null hypothesis. We show that the associated p-value process and confidence intervals (defined as in one-variation experiments) are always valid; see Proposition 7.1 in the Appendix.

For binary data, we consider the one-variation experiment where each observation is a pair (X_n, Y_n) and θ is unknown but μ is fixed. The mixture likelihood ratio in that case reduces to the mixture likelihood ratio based on any sufficient statistic for θ in this one-variation model. We note that for any μ, $\bar{Y}_n - \bar{X}_n$ is asymptotically sufficient with asymptotic distribution $N(\theta, V_n/n)$, where:

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

This distribution resembles that of the sufficient statistic $\bar{Z}_n$ in the normal case with $2\sigma^2 = V_n$, and so by analogy we use the following

mSPRT:

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

where again θ_0 is the difference of success probabilities under the null hypothesis (if $V_n = 0$, we let $\tilde{\Lambda}^{H,\theta_0} = 1$).

Since the approximations hold only at large n, exact always validity is not achieved, but simulations demonstrate that approximate Type I error control is obtained at small α where large sample sizes are necessary to reject H_0. In a companion technical report [9], we provide a more conservative variant of this mSPRT for two-variation binary data, which is proven to control Type I error to leading order as $\alpha \to 0$.

4.4 Implementation details

Across all of Optimizely's products, we implement always valid p-values and confidence intervals using the definitions in (11) and (12) respectively, substituted into the expressions in (9) and (10) respectively.

Some slight modifications are required in the practical implementation. For normal data, we must use a plug-in empirical estimate for σ^2; simulations show that this does not impact the Type I error for small α. Further, there are some continuous-valued metrics such as "$ spend" where the distribution of responses has a heavy right tail, making a normal model inappropriate. For these, a mixture of likelihood ratios are computed under a more general model that can fit this skewness.

We report *statistical significance* at time n; this is $1 - p_n$, shown in Figure 3. Significance begins at zero and increases monotonically, reaching 100% asymptotically if a true difference exists. This has an intuitive benefit: as evidence only increases over time, the level of significance (and hence the user's confidence in inference) should not decrease over time. Similarly, the confidence intervals narrow monotonically.

However, despite its advantages, this monotonicity does present an additional user interface challenge. In a proportion α of tests, the confidence interval will eventually lie entirely above or below the true difference and will never recover, even as the point estimate $\bar{Y}_n - \bar{X}_n$ leaves the confidence interval and approaches the truth. Shifts in the underlying conversion rates during the experiment can amplify this effect.[4] In our deployment, we address this issue with a heuristic "reset policy", which forces a reset in our inference whenever the point estimate leaves the confidence interval: the reported statistical significance is reset to zero at that time, the confidence interval returns to the entire real line, and then the iterations in (9) and (10) begin anew. This policy only ever makes the p-values larger and the confidence intervals wider, so it does not lead to any additional Type I errors. For some choices of stopping time, however, the policy does reduce power.

5 DETECTION PERFORMANCE

Suppose that a user stops the first time that our always valid p-value process crosses α. We know their Type I error is controlled — but what about *detection* of real effects? In this section, we show always valid p-values generated using the mSPRT possess several desirable properties from this perspective. First we discuss several theoretical optimality properties of the mSPRT. Second, we empirically evaluate detection performance under our approach and show that it is often preferable to fixed-horizon testing.

5.1 Optimality

Theoretical results in the literature establish asymptotic optimality properties of mSPRTs. Informally, these results imply that our always valid p-values perform well for users who prioritize high detection over small sample sizes. The most basic such result is that any mSPRT is a test of *power one* [17]; i.e., under any alternative $\theta \neq \theta_0$, the test is eventually guaranteed to reject the null hypothesis, if the user is willing to wait long enough. Further, [12] and [11] establish that the power converges to one quickly as the sample size grows.[5]

But of course, no user can literally wait forever. To formalize this, we suppose that a user's impatience is captured by a *failure time* M, so the user stops the first time the always valid p-value falls below α, or at time M, whichever comes first. This implements the mSPRT truncated to time M. Our companion technical report [9] proves the following key result: *The mSPRT truncated at M achieves significance earlier on average than competing tests that offer the same power, for small α – despite the fact that mSPRT-based p-values are computed* without *knowledge of M.*

In this sense, our always valid p-values provide an "interface" that allows users to nearly optimally trade off detection performance and sample size based on their own preferences.

5.2 Choosing the mixing distribution

The mSPRT as defined in (11) requires a key parameter as input: the mixing variance τ (recall the mixing distribution is $H \sim N(\theta_0, \tau^2)$). Existing theory does not reveal how best to choose this mixture. Intuitively, since Λ_n^{H,θ_0} represents the evidence in favor of a mixture of alternatives $\theta \neq \theta_0$ weighted by $\theta \sim H$, we would expect the best average performance when the mixing distribution H matches the distribution of true effects across the experiments a user runs. Our companion technical report establishes this approximately for normal data and a true prior $G = N(0, \tau_0^2)$ on θ, under appropriate regulatory conditions [9].

At the time of our deployment, customers of Optimizely could purchase subscriptions at one of four tiers: Bronze, Silver, Gold or Platinum. We obtained a prior G on effect sizes separately for each tier by randomly sampling 10,000 two-variation, binary data experiments that had been run previously on Optimizely. The reason for constructing distinct priors across tiers is that customers in higher tiers tended to be further into optimizing their website and so were typically chasing smaller effects. The distribution of effects seen in each tier was indeed approximately normal, and we

[4] Neither fixed-horizon nor always valid measures offer validity guarantees in this case, but the former remain visually reasonable.

[5] For instance, [11] proves the asymptotic Bayes optimality of mSPRTs when the cost to the experimenter is linear in the false positive probability, the power, and the number of observations, and the relative cost of observations approaches zero. For our purposes, given a prior for the true effect $\theta \sim G$ under the alternative H_1, this result implies that if the user prioritizes power and her costs take this simple linear form, our p-values offer her near optimal performance in expectation over the prior, provided she chooses α optimally.

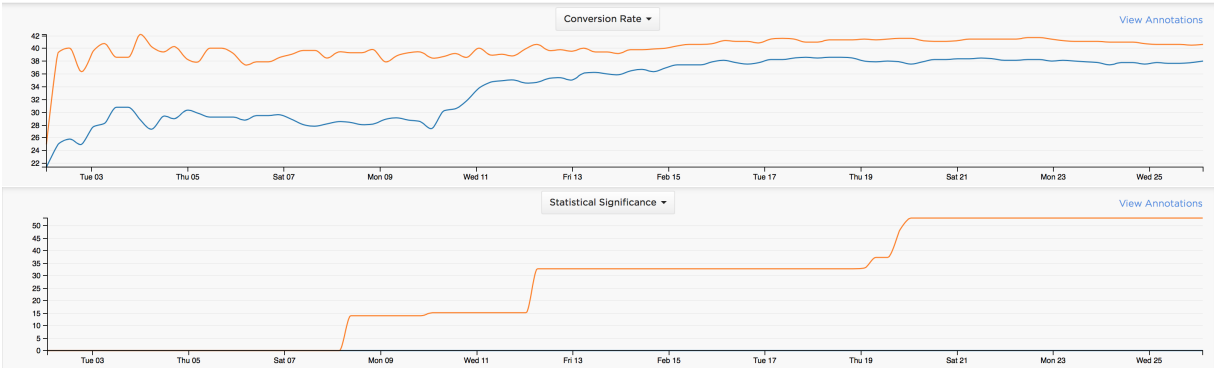

Figure 3: The top chart shows conversion rates over time in an A/B test over time, as displayed in the Optimizely dashboard. The bottom chart shows statistical significance, computed using the always valid p-values described in Section 4.3.

choose τ by fitting a centered normal distribution after appropriate shrinkage via James-Stein estimation. [8]

5.3 Improvement over fixed-horizon testing

We also used the same sample of 10,000 experiments to show that our p-values typically deliver significance faster than a fixed-horizon test, whose sample size is chosen to obtain 80% average power over the prior in each tier. In Figure 4, the red curve shows that in most of the experiments, the mSPRT achieves significance before that sample size.

Of course, the user might choose different fixed-horizons for each experiment if she had additional contextual information that the prior does not capture. The black curves in Figure 4 suppose that the user can estimate the true effect size up to some given relative error, and they compare the expected sample size of the mSPRT in each experiment against the fixed sample size she would choose to achieve 80% power at her estimate. Now we see that fixed-horizon testing outperforms the mSPRT if she can estimate the effect size very accurately. However, since a relative error below 50% is rarely achievable, the mSPRT will typically perform better in practice.

Finally, in Figure 5, we use simulations to evaluate our detection performance for users who may give up for failure at time M (as described above). We consider two-variation experiments with binary data with the true θ drawn from a standard normal prior, and with $\tau = 1$. For four levels of power, β, we choose M so that the average power of the truncated mSPRT equals β, and compare the distribution of sample sizes against the fixed-horizon test achieving that same average power. Since the mSPRT is optimized primarily for users who prioritize detection over sample size, it is outperformed by fixed-horizon testing when β is small. However, for any user who seeks moderate power, the mSPRT p-values generally offer faster detection than fixed-horizon testing.

6 MULTIVARIATE TESTING

Throughout the paper, we have assumed that there are at most two variations under consideration, and only one metric of interest. In fact, when a user initiates an experiment on Optimizely (or most any A/B testing platform), she will often specify many treatment variations that will be compared against a baseline variation, and each visitor in the experiment is randomized between these multiple alternatives. Further, she will often specify several goals on which the variations are to be compared , and each visitor's responses across all of these goals are measured simultaneously. Each comparison defines a two-variation sub-experiment, and the platform displays the results of all sub-experiments a single dashboard (as in Figure 1).

Tests with many goals and variations are called *multivariate tests*. In such a test, when the user attempts to draw inferences across every variation and goal in her experiment, simultaneous false positives in multiple sub-experiments can highly distort her conclusions. As the dashboard in Figure 1 suggests, the user's attention will be drawn to those results which are significant; as the absolute number of false positives increases with the number of goals and variations, the user runs the risk of acting on illusory effects. This is known as the *multiple comparisons* or *multiple hypothesis testing* problem [21].

We conclude the paper by demonstrating how we can combine our always valid measures with methodology from the multiple hypothesis testing literature that protects users from this risk. The resulting Optimizely results page allows users to continuously monitor tests with many variations and goals, and yet remain confident in the inferences drawn.

There are two well-studied methods in the multiple testing literature for testing an arbitrary number m of hypotheses using m correlated data sets (note that observations across our sub-experiments are correlated): the *Bonferroni correction* and the *Benjamini-Hochberg (BH) procedure*. The Bonferroni procedure is designed to control the *family-wise error rate* (FWER): the probability that any true nulls are rejected [4]. The Benjamini-Hochberg procedure controls the *false discovery rate* (FDR): the expected proportion of rejected null hypotheses that are in fact true [2].

Each procedure operates as follows. The inputs are the computed p-values for the m hypothesis tests, $p_1, \ldots, p_m$. The outputs are "corrected" values, one per hypothesis, that we refer to as *q-values*: $q_1, \ldots, q_m$. The policy of rejecting the null hypothesis in test j if $q_j \leq \alpha$ controls the FWER at level α (using the Bonferroni q-values)

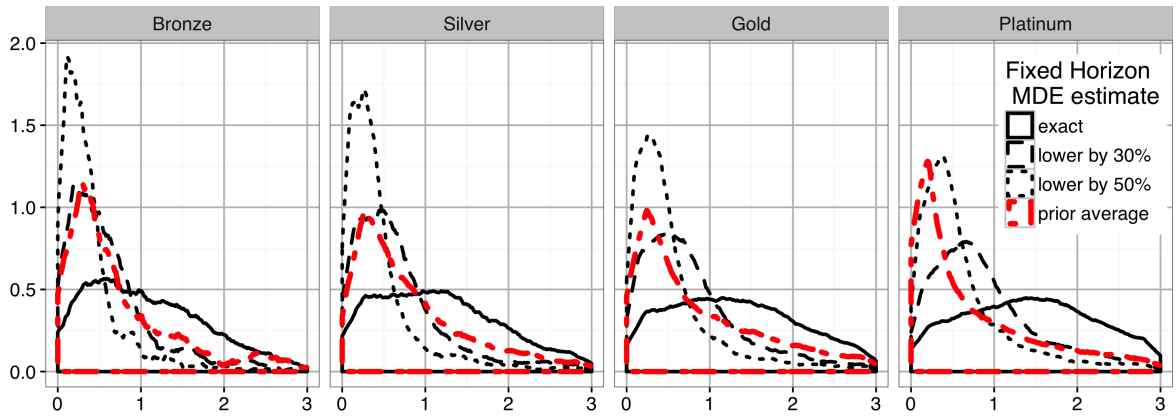

Figure 4: The empirical distribution of sample size ratios between the mSPRT and fixed-horizon testing over 10,000 randomly selected experiments, divided up by subscription tier.

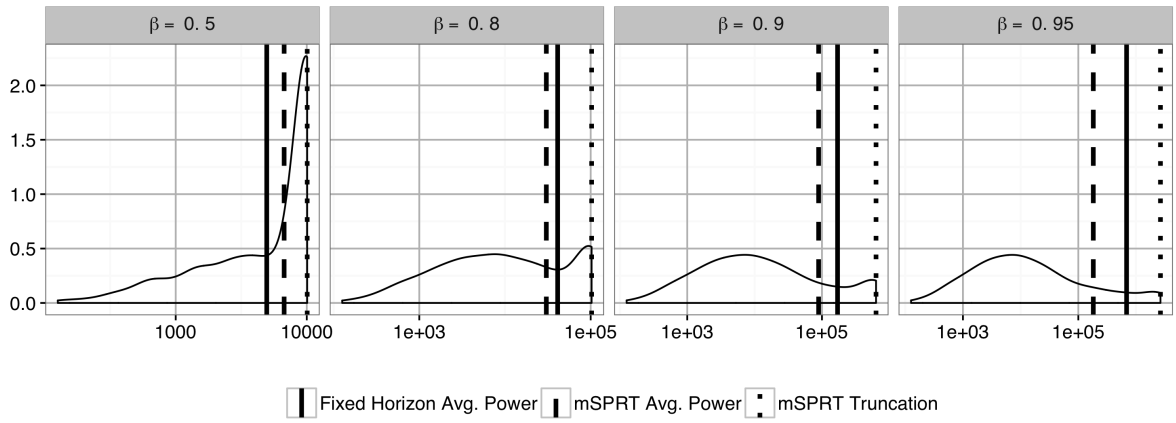

Figure 5: The simulated distribution of sample sizes for the mSPRT truncated to give four different average powers. The average run-time (dashed) and truncation sample size (dotted) are compared against the fixed-horizon (solid) that achieves the same average power.

and controls the FDR at level α (using the BH q-values). The exact formulae for these computations are given in the Appendix.

A great virtue of the fact that our approach to continuous monitoring employs p-values is that we can directly leverage these methods for multiple testing corrections as well. For either method, the computations are straightforward; we display q-values on Optimizely's dashboard computed directly from the mSPRT-derived always valid p-values. Monotonicity of the p-values ensures monotonicity of the q-values under either the Bonferroni or BH procedure. Most importantly, it is straightforward to demonstrate that the Bonferroni or BH q-values obtained from any collection of always valid p-values control FWER or FDR (respectively) in the presence of arbitrary continuous monitoring.

In general, there are situations in A/B testing practice where Bonferroni q-values, BH q-values, or even uncorrected p-values may help users to make decisions most effectively. The dilemma is that FWER control provides the safest inference, but Bonferroni offers less detection power at any given sample size than BH, which itself reduces power compared with no correction. From user research, we decided on FDR control as it appeared to best reflect how Optimizely's customers intuited their results when making decisions: they focused mostly on the significant results displayed on the dashboard and expected most (but not all) of these to be accurate.[6]

[6]In some cases, however, users include many goals but make decisions primarily on one goal. We additionally allow to the user to select a "primary goal" and provide FDR

REFERENCES

[1] Akshay Balsubramani and Aaditya Ramdas. 2015. Sequential nonparametric testing with the law of the iterated logarithm. *arXiv preprint arXiv:1506.03486* (2015).

[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* (1995), 289–300.

[3] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* (2001), 1165–1188.

[4] Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.

[5] Ronald Aylmer Fisher and others. 1949. The design of experiments. *The design of experiments.* Ed. 5 (1949).

[6] Bhaskar Kumar Ghosh and Pranab Kumar Sen. 1991. *Handbook of sequential analysis.* CRC Press.

[7] John M Hoenig and Dennis M Heisey. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 55, 1 (2001), 19–24.

[8] William James and Charles Stein. 1961. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1. 361–379.

[9] Ramesh Johari, Leo Pekelis, and David J. Walsh. 2015. Always Valid Inference: Bringing Sequential Analysis to A/B Testing. (2015). arXiv:arXiv:1512.04922

[10] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 1168–1176.

[11] Tze Leung Lai. 1997. On optimal stopping problems in sequential hypothesis testing. *Statistica Sinica* 7, 1 (1997), 33–51.

[12] T. L. Lai and D. Siegmund. 1979. A Nonlinear Renewal Theory with Applications to Sequential Analysis II. *The Annals of Statistics* 7, 1 (01 1979), 60–76. DOI: http://dx.doi.org/10.1214/aos/1176344555

[13] Erich Leo Lehmann, Joseph P Romano, and George Casella. 1986. *Testing statistical hypotheses.* Vol. 150. Wiley New York et al.

[14] Evan Miller. 2010. How not to run an A/B test. (2010). http://www.evanmiller.org/how-not-to-run-an-ab-test.html Blog post.

[15] Evan Miller. 2015. Simple Sequential A/B Testing. (2015). http://www.evanmiller.org/sequential-ab-testing.html Blog post.

[16] Herbert Robbins. 1970. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* (1970), 1397–1409.

[17] H Robbins and D Siegmund. 1974. The expected sample size of some tests of power one. *The Annals of Statistics* (1974), 415–436.

[18] 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.

[19] David Siegmund. 1985. *Sequential analysis: tests and confidence intervals.* Springer.

[20] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 17–26.

[21] John W Tukey. 1991. The philosophy of multiple comparisons. *Statistical science* (1991), 100–116.

[22] Abraham Wald. 1945. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* 16, 2 (1945), 117–186.

control on that goal in isolation. BH q-values are then computed for all other goals together.

7 APPENDIX

Proof of Theorem 3.2. Let T be a stopping time. Nestedness of the sequential tests implies that, for any $s \in [0,1], \varepsilon > 0$:

$$\{p_T \leq s\} \subset \{T(s+\varepsilon) \leq T, \delta(s+\varepsilon) = 1\} \subset \{\delta(s+\varepsilon) = 1\}.$$

$\therefore \mathbb{P}_{\theta_0}(p_T \leq s) \leq \mathbb{P}_{\theta_0}(\delta(s+\varepsilon) = 1) \leq s + \varepsilon$, and so the result follows on letting $\varepsilon \to 0$. For the converse, it is immediate from the definition that the tests are nested. For any $\varepsilon > 0$

$$\mathbb{P}_{\theta_0}(\delta(\alpha) = 1) = \mathbb{P}_{\theta_0}(T(\alpha) < \infty) \leq \mathbb{P}_{\theta_0}(p_{T(\alpha)} \leq \alpha + \varepsilon) \leq \alpha + \varepsilon$$

where the last inequality follows from the definition of always validity. Again the result follows on letting $\varepsilon \to 0$. □

Proposition 7.1. *For normal data, the p-value and confidence interval associated with this two-sample mSPRT are always valid.*

Proof. Given any μ^A, μ^B with $\theta = \mu^B - \mu^A$, the distributions of these p-value and confidence interval processes under $\mathbb{P}_{\mu^A,\mu^B}$ are equal to the distributions of the one-sample mSPRT p-value and confidence interval processes based on (Z_n) under $\mathbb{P}_\theta$. □

7.1 Computing q-values

The formulae to compute Bonferroni or BH q-values from current p-values $p^1, \ldots, p^m$ are as follows.

For the Bonferroni correction, we define $q^i = \min\{mp^i, 1\}$ for each hypothesis i.

For the BH procedure, we derive the q-values as follows. Let $p^{(1)} \leq \cdots \leq p^{(m)}$ denote the p-values placed in ascending order. Let $q^{(m)} = p^{(m)}$, and for $i = m-1, \ldots, 1$, define:

$$q^{(i)} = \min\left\{\frac{m(\sum_{j=1}^{i} 1/j)\, p^{(i)}}{i}, q^{(i+1)}\right\}.$$

Note that we include the term $\sum_{j=1}^{i} 1/j$ to account for the fact that the p-values may be correlated (at the very least, since the user's stopping time induces some correlation). See [3] for details.