The New Stats Engine

Leo Pekelis*[1,2], David Walsh†[1,2], and Ramesh Johari‡[1,3]

[1]Optimizely
[2]Department of Statistics, Stanford University
[3]Department of Management Science & Engineering, Stanford University

Abstract

Optimizely is moving away from traditional, fixed horizon hypothesis testing to sequential testing and replacing Type I error control with false discovery rate (FDR) control. This will enable users to confidently check experiments as often as they like, not need to know an effect size in advance, and test as many variations and goals as desired without worrying about hidden sources of error. Read on to understand why, how, and what new tradeoffs exist.

1 Introduction

A Type I error, or false positive, occurs when a statistical decision procedure rejects the null hypothesis, even though the null hypothesis was true, i.e., where there is no difference between the treatment and control groups. The Type I error *rate* is the frequency with which such errors are made on true null hypotheses; in other words, how often do we declare a result significant on an A/A test? For example, if our procedure has a 10% Type 1 error rate, then we expect about 1 detection in 10 A/A tests, 2 in 20, 5 in 50, and so on. Type I error rate is a key statistical measurement, and indeed it is the underpinning of any valid statistical decision procedure.

Type I error is the "yin" to the "yang" of statistical power, which is the probability that an A/B test with a true difference between treatment and control will be declared significant, and called in the right direction. Since decision procedures based on data all have underlying uncertainty, requiring zero chance of a Type I error, also means zero power because no tests can ever be called significant. Conversely, having 100% power also means 100% Type I error, since all tests must lead to significant detections. In general, the experimenter trades off one against the other: increasing

*leonid@optimizely.com
†david.walsh@optimizely.com
‡ramesh.johari@stanford.edu

power brings with it increasing Type I error rate, and "good" statistical procedures are those that yield relatively high power while maintaining a reasonably low Type I error rate. For example, most A/B testing platforms suggest controlling Type I error rate at around 5-10%, and choosing a sample size to ensure statistical power is at least 80%.

Optimizely's new Stats Engine is designed to better match the way experimenters use the platform, while still providing both control of statistical error and good power. (See this post for more details on the vision behind the new Stats Engine.) In this article, we detail some of the technical changes under the hood, focusing on two innovations in particular: sequential hypothesis testing, and control of false discovery rate for multiple hypothesis testing.

Section 2 introduces continuous monitoring (2.1), discusses why continuous monitoring inflates Type I error under classical, fixed horizon statistics (2.2), presents our formulation of sequential testing (2.3), and how to calculate significance from a sequential test (2.4). Section 3 introduces the multiple testing problem (3.1), false discovery rates (3.2), why they provide key actionable insights missing from Type I error control (3.3), why they work (3.4), and how they extend to confidence intervals (3.5) and commute with sequential testing (3.6). Section 4 contains a technical appendix pertaining to section 3. Finally, section 5 is a reference appendix, connecting together all our results.

2 Continuous monitoring and sequential testing

We refer to the traditional statistics used in A/B testing as fixed-horizon statistics; it presumes that you must set your sample size in advance. In most historical areas of experimentation, this has been an unavoidable logistical constraint; e.g., in agricultural studies you must decide the quantity of crops to plant in advance, with no data to guide you until all your crops have grown. In such contexts, fixed-horizon statistics is optimal in the following sense: given a bound on the proportion of false positives, the probability of detecting any true effect with the predetermined data set is maximized, irrespective of the size of that effect.

For modern A/B testing the logistical constraints fall away, and with the detailed real-time data we have now it should be possible to do much better. One of the greatest benefits of advances in information technology, computational power, and visualization is precisely the fact that experimenters can watch experiments in progress, with greater granularity and insight over time than ever before. A key goal of the new Stats Engine is to take advantage of this opportunity with a novel statistical approach.

How can we take advantage of real-time data? Note that there is a subtlety in the optimality property of fixed-horizon statistics described above: while the power (probability of detection) is maximized for any true effect, the power can still be very low if the effect is small compared with the sample size. To have adequate power for detecting small effects, the experimenter will have to commit to a large sample size up front, and this will be wasted effort if it turns out the true effect could have reliably been detected much sooner.

When a sample size need not be set in advance, the data should let you learn the size of the effects to consider adaptively, and so optimize the sample size at which the test terminates. But fixed-horizon statistics offers no mechanism for mid-experiment optimization, so unfortunately the additional flexibility of A/B testing has so far only been considered detrimental, driving up the number of false positives if you are not savvy.

In this section, we explain how continuous monitoring of an experiment increases the Type I error rate when fixed-horizon statistics is used, and we set out our solution that leverages real-time data correctly through a framework known as Sequential Testing.

2.1 Statistical decision procedures, continuous monitoring, and decision boundaries

To begin, let's think about how we reach a conclusive result. For convenience we will just consider a binary goal, although the discussion is essentially the same for real-valued goals. Let's assume we have just the baseline (the "control") and a single variation (the "treatment"). We'll use n to index the number of visitors that have arrived. To keep things simple, let's assume the allocation is evenly split so that $n/2$ visitors are in the baseline and variation; all the discussion below is equally valid (just with more notation) even if the allocation is unequal. For the rest of the presentation, we'll use $\bar{X}_n, \bar{Y}_n$ to represent the current observed conversion rate for the baseline and the variation respectively. Finally, the key quantity of interest to us is the *difference* of these conversion rates, $\hat{\theta}_n = \bar{Y}_n - \bar{X}_n$.

Experimenters run A/B tests because they want to learn which variations are winners (and which are losers). Intuitively, this means we should declare results conclusive if $\hat{\theta}_n$ is "significantly" different from zero (either positive or negative). In other words, we can think of a statistical decision procedure as telling us how large the magnitude of $\hat{\theta}_n$ needs to be for results to be conclusive. A classical result in statistics (the Sufficiency Principle) tells us that all good decision procedures have this form.

To see this approach in action, let's revisit fixed-horizon statistics. If you had committed to a sample size of n, then fixed-horizon statistics would tell you that the test is conclusive at a given level α (e.g. 5%) if

$$|\hat{\theta}_n| > k\sqrt{V_n}, \tag{1}$$

where:

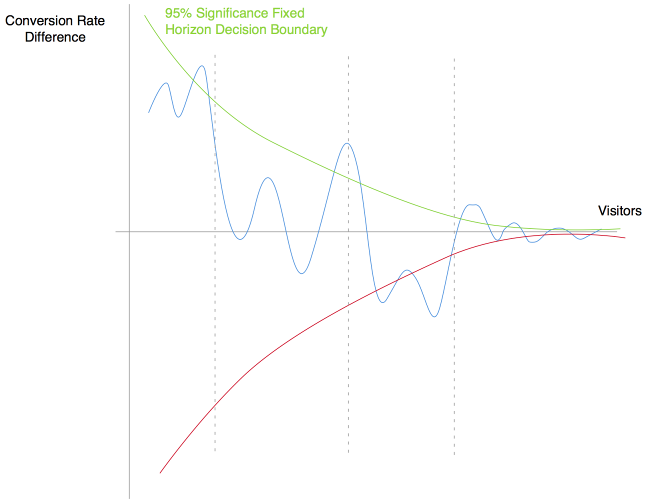
$$V_n = \frac{2(\bar{X}_n(1-\bar{X}_n) + \bar{Y}_n(1-\bar{Y}_n))}{n}.$$

and k is a fixed constant.

Note that for binary goals, V_n is an estimate of the variance of the observed difference of conversion rates; in particular, it is the sum of the estimate of the variances of $\bar{X}_n$ and $\bar{Y}_n$, respectively. Given a desired control α on Type I error probability, the constant k can be chosen to ensure that the Type I error is equal to α. There is a fundamental tradeoff here: if k is high, we are less likely to declare a result significant; this means we have a lower rate of Type I error (false positives), but at the expense of lower statistical power (detection of true effects).

Of course, as noted in the blog post here, fixed-horizon statistics ensures that Type I error is controlled under the assumption that *the experimenter only looks at the data once.* However, if rather than setting the sample size in advance, the experiment is continuously monitored as it is running, there are many potential decision points. When viewing a test through the lens of continuous monitoring, it's useful to visualize the *decision boundary* of a procedure. For example, if a test is stopped the *first* time that (1) is satisfied, the decision boundary as a function of n is the right hand side of (1). We've plotted this decision boundary in Figure 1.

Figure 1: Examples of two 95% Fixed Horizon decision boundaries (solid curves). The dashed lines represent looking at the data more than once.



2.2 Why does continuous monitoring inflate Type I error?

A key point that we emphasize is that continuous monitoring of fixed-horizon statistics severely inflates Type I errors, losing signal about truly significant results in a sea of false positives. In this section we walk through an example to see why this happens.

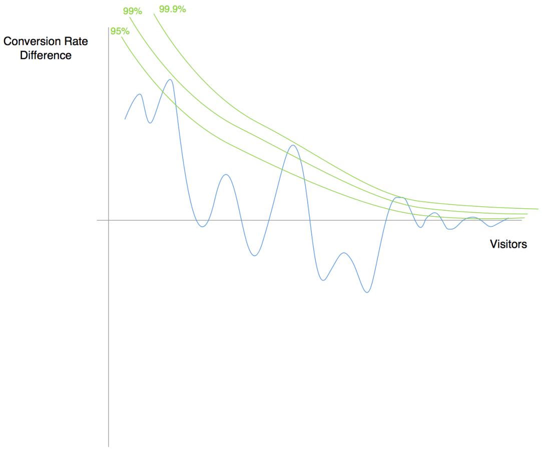
Suppose that the experimenter were to stop the first time that the (1) is satisfied. What should we expect to see? Such a rule is *guaranteed* to report a false positive; in other words, the test will eventually cross the decision boundary under the null hypothesis of no difference between baseline and variation. To see why, note that as $n \to \infty$, the right-hand side of (1) is approximately proportional to $1/\sqrt{n}$. However, it can be shown that under the null hypothesis, the largest fluctuations of $\hat{\theta}_n$ are approximately proportional to $\sqrt{(\log \log n)/n}$; this is known as “ the law of the iterated logarithm” . Precisely, under the null hypothesis, this law implies that with probability one, there exist arbitrarily large values of n for which

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

The right hand side is guaranteed to eventually exceed any threshold that is proportional to $1/\sqrt{n}$; in particular, it is guaranteed to exceed the decision boundary on the right hand side of (1), so the Type I error rate of such a rule is 100%!

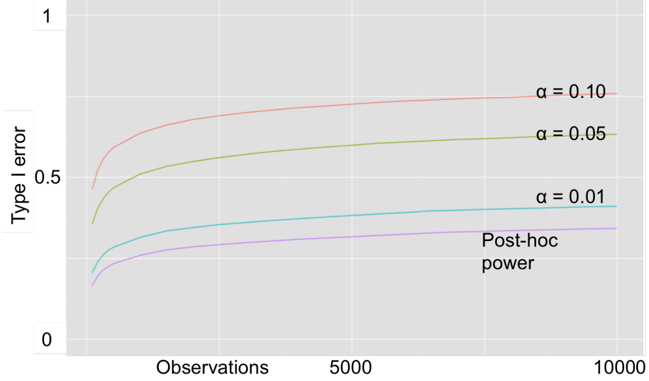
In practice, the Type I error of such a rule is somewhat lower because the experimenter doesn't wait indefinitely. Even so, with reasonable sample sizes, it is typically the case that Type I error rate will exceed 70-80%. So what is the experimenter to do? It's clear from the preceding decision that becoming more conservative (i.e., using a higher constant k in (1)) doesn't work, because regardless of the value of k the law of the iterated logarithm says the decision boundary will eventually be crossed. (See Figure 2.)

Figure 2: More conservative fixed horizon boundaries still have high Type 1 Error



In our own conversations with our customers, we've heard of a range of approaches that are taken to try to combat the inflation of Type I errors that continuous monitoring can introduce. A common approach used is to wait until the fixed-horizon test is conclusive, *and* a sample size calculation using the currently observed $\hat{\theta}_n$ as the effect size says that the test has run “long enough”, i.e., that the current n is sufficient to achieve a given power (e.g. 80%). This is called a *post-hoc power* approach. Surprisingly, this strategy can be shown to be mathematically equivalent to choosing a more conservative k in (1)! (See [3] for discussion.) As such it is also guaranteed to be eventually crossed. Being more conservative with a larger k can be useful, but practically relevant values of k at commonly observed sample sizes still lead to Type I error rates of 30-40%, well above the 5% or 10% levels that experimenters usually set. (See Figure 3.)

Figure 3: This figure shows how Type I error rate scales at different levels of desired Type I error control α, as well with a post-hoc power approach with a 5% α level.



2.3 Towards a solution

Our goal is to let the experimenter make decisions as soon as possible. However, naive continuous monitoring can lead to spurious decisions - ones based on chance trends in experimental data - overwhelming good ones.

The previous section illustrates the problem: while a decision boundary that scales like $1/\sqrt{n}$ works when we look only at a single point in time, it does not prevent the sample path of $\hat{\theta}_n$ under the null hypothesis from crossing at *any* point in time. This is what causes the spike in Type I error with continuous monitoring: eventually, $\hat{\theta}_n$ will cross the decision boundary. What is needed is a decision boundary that ensures the probability that $\hat{\theta}_n$ *ever* crosses is controlled at a desired level α.

One simple way of choosing such a boundary is to choose a very large constant C, and only stop if $|\hat{\theta_n}| > C$. This will indeed lead to a very low Type I error, since the chance of $|\hat{\theta}_n|$ crossing a fixed large value under the null hypothesis can be made arbitrarily low. The problem, of course, is that this probability will *also* be low even if there really is a difference — in other words, the statistical power of such a boundary will be terrible. What we want is to ensure that we will detect an effect if one really exists. Note that if the true conversion rate difference is non-zero, $\hat{\theta}_n$ will converge to this true value, so we need a decision boundary that converges to zero as $n \to \infty$. Such a boundary is guaranteed to be crossed when the effect is nonzero; in fact, the faster such a boundary converges to zero, the faster we will detect true effects — at the expense of a potentially inflated Type I error rate.

So this is the problem before us: Can we find a decision boundary that converges to zero quickly (so that true effects are detected fast), but is not crossed too often under the null hypothesis (so that Type I error is controlled)? A clue to the solution can be found in the discussion of the previous section, from the law of the iterated logarithm. We know that $|\hat{\theta_n}|$ will cross $\sqrt{(\log \log n)/n}$ infinitely often, so we need a boundary that goes to zero slower than this. However, we want a boundary that goes to zero as fast as possible.

Our solution, inspired by this observation, uses a boundary proportional to $\sqrt{\log n/n}$. This boundary goes to zero as $n \to \infty$, so that all true effects are detected. But it doesn't do so overly quickly, so that Type I error can be controlled.

Formally, suppose that the experimenter has fixed a desired bound α on Type I error. With Stats Engine, the test is deemed conclusive once the following condition holds:

$$|\hat{\theta}_n| > \sqrt{\left(2\log\left(\frac{1}{\alpha}\right) - \log\left(\frac{V_n}{V_n+\tau}\right)\right)\left(\frac{V_n(V_n+\tau)}{\tau}\right)}. \tag{2}$$

First, note that V_n scales approximately proportionally to $1/n$ for large n. Using this fact and a little algebra, it is straightforward to check that the right hand side is approximately proportional to $\sqrt{\log n/n}$ for large n.

Second, note that there is a constant τ that determines the exact decision boundary. This constant is a free parameter, in the sense that regardless of the value of τ (as long as it is positive), the resulting test gives valid Type I error control while ensuring any true nonzero effect will eventually be detected. However, careful tuning of τ can have a significant effect on the speed of detection of true effects.

Where does (2) come from? The Technical Appendix below details the derivation further, but the basic idea for the boundary can be summarized in a few steps. The key to our analysis is to consider a *likelihood ratio* test. In particular, fix a positive θ; the likelihood ratio (LR) given the data so far is the ratio of the probability of seeing the data if the true difference were θ, to the probability of seeing the data if the true difference were zero. When the LR is large, it means that the true effect of θ is more likely than the null hypothesis of zero difference. As a result, we can consider a simple rule: *stop if the LR ever crosses a fixed threshold.* Prior results in sequential testing (see [7, 5, 6, 4]) have shown that this test (with an appropriate choice of fixed threshold) has the desired properties, *if* we knew in advance the only possible value of a true effect is θ. This test is known as the Sequential Probability Ratio Test (SPRT).

Unfortunately, in practice, we don't know the true effect θ. Of course, we do have lots of prior knowledge about what kinds of effects might be "reasonable" to see, and so a natural approach in this case is to think like a Bayesian: put a prior distribution on θ. Formally, we can compute the *average LR* against a prior distribution on the true effect θ, and wait to see if this average LR ever crosses a fixed threshold. (See [5, 6]). This test (rewritten) is exactly the test described in (2). A schematic representation is given in figure 4.

The prior lets us focus on the effect sizes we most anticipate, and the test is optimized for fast detection on θ where the prior is large, i.e., for those effects we are likely to observe. For the normal prior, this is controlled by the parameter τ. We chose our prior, and hence the parameter τ, as the result of extensive analysis of historical experiments run on Optimizely's platform. It should be noted that without this personalization, sequential testing did not give results quickly enough to be a viable for use in an industry platform. This is something we intend to keep refining going forward.

2.4 Statistical significance

As described in our first blog post on the new Stats Engine, one of our key goals is to ensure users can easily interpret the significance level of a test. Users shouldn't have to carry out any secondary diagnostics to control for the effects of continuous monitoring. To give users this ability, we need to generate significance levels from the decision procedure described above.

In classical statistics, the significance level is closely related to the *p-value*. Informally, the p-value reports the chance of seeing the observed data, if the null hypothesis were true; the significance level is one minus the p-value. In other words, a high significance level suggests it is safe to declare a test as conclusive.

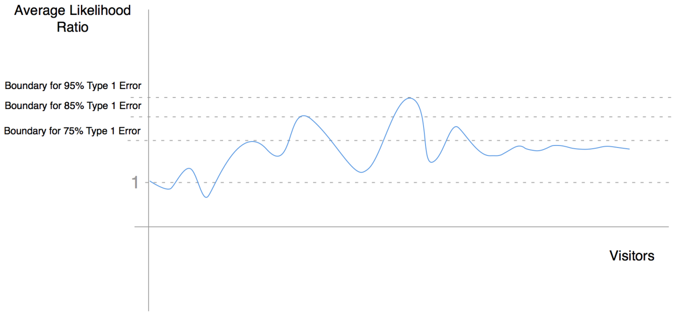
There is a straightforward way to obtain p-values from a decision rule like the one in (2). In particular, we define the p-value after n visitors to be *the least α such that the threshold has already been crossed*. We already know that under the null hypothesis, the chance of ever crossing the threshold in (2) is α. With our definition, this property guarantees that the chance the p-value ever drops below α is no larger than α.

What does this mean in practice? Regardless of when you choose to look at the experiment, you are guaranteed that if you call the test when the p-value is below α (i.e., your statistical significance is above $1 - \alpha$), your Type I error rate will be no more than α. We achieve this strong control on false positives, while still making it possible to call tests as quickly as possible when the data justifies it (giving good statistical power). This balancing of dual goals is what makes sequential testing so powerful.

3 Multiple hypothesis testing and controlling false discovery rate

In this section, we tackle the second key challenge of the new Stats Engine: making it easier for customers to include multiple goals and variations, and confidently interpret the results. While controlling Type I error is a fundamental aspect of nearly all statistical decision procedures, there

Figure 4: A Sequential test monitors the Average Likelihood Ratio to draw decision boundaries at different levels of significance

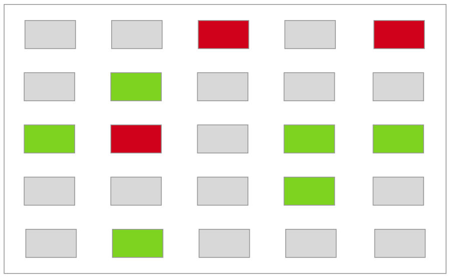


is a significant pitfall to be aware of as the number of simultaneous tests increases: controlling Type I error rate does not lead to actionable insight as the number of tests increases.

3.1 The multiple hypothesis testing problem

As an example, suppose the matrix with boxes in Figure 5 represents a set of A/B tests. Boxes with no color denote inconclusive outcomes, while green and red boxes denote conclusive winners and losers. In other words, a box is the outcome of one variation on a goal.

Figure 5: A set of A/B tests. Each box represent a goal, variation combination



A Type I error guarantee, with $\alpha = .1$, on this matrix translates to the following statement: "Not more than 10% of boxes would be called winners or losers, if all the boxes corresponded to A/A tests."

The problem for the decision maker is that the guarantee on Type I error is not actionable on this matrix. If there were a lot more reds and greens than 10% of the boxes, we might conclude that some of the tests have a true difference between variation and baseline. If less than 10% were conclusive, we would think that the whole matrix is filled with inconclusive tests. But how many is "some of the tests"? What is the relationship with "a lot more than 10%"? And most importantly, which A/B tests can the decision maker take away as having a significant impact on her business?

There is no way to draw these conclusions from the guarantee that Type I error has been controlled.

In the remainder of this post we discuss an alternative quantity that can be controlled when multiple tests are simultaneously run: the false discovery rate. The key benefit of false discovery rate control is that it leads more directly to actionable business insight in the presence of multiple hypothesis tests.

3.2 What are false discovery rates?

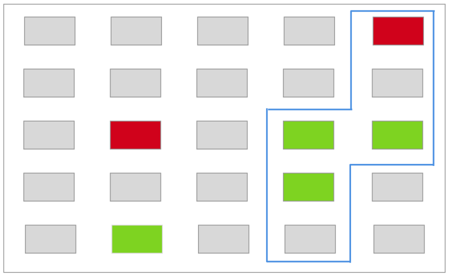
False discovery rate (FDR) is a different measurement designed to address the issues with reporting Type 1 error while running many concurrent experiments. It is defined as the expected proportion of false detections among all detections made:

$$\mathsf{FDR} = E\left[\frac{V}{R \vee 1}\right], \tag{3}$$

where V is the number of tests declared significant which were truly null (i.e., effectively like A/A tests), and R counts the overall number of tests declared significant. The convention is to take $V/R = 0$ if no rejections were made ($R = 0$), as denoted by the V sign for maximum.

Consider the matrix of A/B tests again. If the blue outline in figure 6 denotes the variations which are truly different from the baseline, then the proportion of false discoveries are the number of red and green boxes falling outside the blue area, divided by the total number of red and green boxes.

Figure 6: False discovery rate controls proportion of winning and losing declarations that are incorrect



Benjamini and Hochberg showed that any hypothesis testing procedure designed to control Type 1 error rate can be transformed into a procedure for controlling false discovery rate using a straightforward method [1]. It is this method which underlies our approach to multiple testing at Optimizely.

3.3 Does controlling the false discovery rate lead to actionable insight?

To get a hint at why FDR is useful in a business decision making context, let's start by comparing how the decision maker might react to seeing, say, 5 conclusive tests (winners and losers combined). If all that is known is that the decision procedure controlled Type I error, then the number of other tests run is very important: if 50 other tests were run, these 5 may be highly likely to be conclusive by random chance alone; but if only 5 other tests were run, then the decision maker may become more confident the 5 conclusive results are meaningful. In other words, the number of other inconclusive tests affects whether the decision maker believes her results are actionable!

This perverse consequence is eliminated with procedures that control FDR. Reporting a set of A/B Tests with, for example, "false discovery rate less than 10%" is completely actionable without knowing anything about non-significant variations. The statement means that at least 90% of the reported red and green boxes have a true difference between variation and baseline, regardless of how many remaining boxes are in the matrix. By themselves, they are a very promising group of variations to act on in order to optimize your business.

Let's dig a little deeper and try to formalize why FDR estimates are more actionable with multiple A/B tests. Mathematically, the main observation is that FDR reverses the conditional probability that defines Type I error.

In classical statistics, a p-value reports estimates of *the probability of the data given there is no difference between baseline and variation.* False discovery rates reverse this conditional probability: in particular, FDR is an estimate of *the probability of no difference (or a false detection) between the baseline and variation, given the data you observed.*

The questions from before are exactly answered by only reporting conclusive results in those boxes where FDR is less than a threshold, say 10%. Note that $1 - \mathsf{FDR}$ is an estimate of the chance there is a true difference between baseline and variation, given the data observed.

In fact, it is possible to further put an ordering on the significance level of significant A/B tests. If we have a group of conclusive boxes when we control FDR at 10%, and a (smaller) conclusive subset of boxes when we control FDR at 1%, we can say that the whole group of significant boxes is 90% likely to give you actual lift when implemented, while the smaller group is 99% likely. Risk trade-offs guide a consumer of these results to make judgements on which variations to implement or further investigate.

It is for these reasons that we at Optimizely now report significance as $1 - \mathsf{FDR}$ for each A/B test.

3.4 Why does FDR control work?

In this section we formalize the relationship between FDR control and the conditional probability

$$P(\theta = 0 \mid \hat{\theta}_n)$$

using the notation from section 2.1. In this section, we overload $\hat{\theta}_n$ slightly and use it as shorthand for "all observed data required for the statistical decision procedure at hand." In particular, it may now also depend on quantities like V_n. We can use the definition of conditional probability to write

$$P(\theta = 0 \mid \hat{\theta}_n) = \frac{P(\theta = 0 \cap \hat{\theta}_n)}{P(\hat{\theta}_n)} = \frac{P(\theta = 0)P(\hat{\theta}_n \mid \theta = 0)}{P(\hat{\theta}_n)}$$

A (empirical) Bayesian perspective on false discovery rate control is that we try to estimate the three quantities on the right of the equals sign to get an estimate of the conditional probability on the left. Then we report all hypotheses with an estimate lower than a pre-specified threshold.

There are three steps. First, notice that $\hat{p} \equiv P(\hat{\theta}_n \mid \theta = 0)$ is simply the Type 1 error, or p-value, and we have those for each A/B test. Next assume that there is an unobserved prior proportion, π_0, of null tests (no true difference).

Finally, you may have noticed that we have used the words data, observed difference, and p-value interchangeably. This is because a p-value is the amount of evidence collected against a null hypothesis of no true difference in conversion rates ($\theta = 0$). Smaller p-values denote more evidence against the null hypothesis so

$$P(\hat{\theta}_n) = P(p \leq \hat{p})$$

where $\hat{p} = \hat{p}(\hat{\theta}_n)$ is the p-value as a function of $\hat{\theta}_n$, and this probability is computed over all tests, whether $\theta = 0$ or not. The unbiased, maximum likelihood estimate of this probability is just the number of boxes with p-values as small as the p-value of the test, divided by the number of boxes in the matrix (N). Or, if we order the A/B tests by their p-values, from smallest to largest, then

$$P(\theta = 0 \mid \hat{\theta}_{n,i}) = P(\theta = 0 \mid \hat{p}_i) = \frac{\pi_0 \hat{p}_i}{i/N} \tag{4}$$

is the estimated FDR of the ith A/B test. The Benjamini-Hochberg method upper bounds $\pi_0 \leq 1$, and then rejects hypotheses based on this estimate. Since the Benjamini-Hochberg method controls the FDR described in (3), we get a duality between FDR and $P(\theta = 0 \mid \hat{\theta}_n)$.

The result is a group of hypotheses, all of which are guaranteed to have both definitions of false discovery rate, (3) and (4), below a threshold. The winning and losing variations (green and red boxes) Optimizely reports are those with FDR less than the value set in your project level settings. The significance value Optimizely reports for each A/B tests are estimates of $P(\theta \neq 0 \mid \hat{p}_i)$.

3.5 Using FDR for confidence intervals

Along with false discovery rate is a closely related measurement which is called false coverage rate (FCR). False coverage happens when a confidence interval (on the lift in conversion rates) does not contain the true difference between variation and baseline, θ.

Intuitively, one wants a method controlling FCR to give the same guarantee as one for FDR, for example, at most 10% of significant confidence intervals do not contain the true difference. This statement seems innocuous. By definition, a marginal confidence interval on one A/B test already attains at most 10% chance of not covering the true difference, so looking at a collection of confidence intervals should similarly have at most 10% error rate of coverage. But as with false discovery rate, the danger lies in the single word *significant* because it encapsulates a selection procedure. The confidence intervals that end up being reported are not on a random selection of A/B tests, but on a specific subset of A/B tests - the ones that had high significance values. This selective action by the experimenter changes the coverage rate.

As a simple example, suppose we use the selection procedure - "only report a confidence interval if it doesn't contain 0." Then no matter what coverage level the marginal confidence interval had, it now has exactly 0 chance of covering the true difference between variation and baseline when it is 0. This phenomena was first examined in [2], who defined FCR as

> the expected proportion of parameters not covered by their confidence intervals among the selected parameters

and gave a procedure for controlling this quantity. We describe this procedure below and modify it slightly to be applicable in an A/B testing platform.

One procedure that guarantees FCR control regardless of the selection procedure is to require that at most 10% of *all* confidence intervals do not contain the true difference, and is achieved by reporting $1 - q/N$ confidence intervals where q is the desired error rate and N the number of A/B tests. This is too stringent for the same reason that thresholding the chance of at most 1 false detection (one red/green box in the non-blue area of figure 2) is. If we do this, then the FCR is controlled too well in the sense that FCR is closer to 0 than q and so confidence intervals are too wide.

On the other hand, FCR should be controlled at the very least on all winners and losers. This led [2] to define a FCR controlling procedure which reports $1 - qm/N$ confidence intervals (m is the number of significant variations). This lower bounds the proportion of false coverage by $q/2$ and has confidence intervals which do not contain 0 exactly when they are winners and losers by the FDR thresholding procedure described in section 3.4.

We argue that extending FCR to an A/B testing platform means that FCR should also be controlled on the test currently visible on a user's dashboard, regardless of its significance. The act of viewing an A/B test must mean that it is a candidate in an experimenter's selection procedure. Similarly, unobserved and non-significant A/B tests cannot be because we stipulate that a current significance level must be observed for selection.

To this end, we implement the following procedure for generating confidence intervals with FCR controlled at level q: for the ith box, report a standard confidence interval using the following altered nominal coverage levels

- if the box is significant, set coverage level to $1 - q\frac{m}{N}$
- if the box is not significant, set coverage level to $1 - q\frac{m+1}{N}$.

where: m is the number of conclusive results after applying a method to control FDR at threshold q.

This procedure has FCR control on precisely the confidence intervals of the set of boxes with estimated FDR below a threshold, and, in addition, the confidence interval currently viewed by the user, regardless of its significance.

Our modification shows another fundamental difference between online experimentation and classical hypothesis testing. An experimenters behavior may be directly observed during the test, and this observed behavior can be directly incorporated into inferential guarantees.

3.6 Does FDR control work with sequential testing?

We have designed our approach to sequential testing to commute with applying the Benjamini-Hochberg procedure at every point in time, and still give the expected guarantees. In other words, if we define FDR_t as the expected false discovery rate after t minutes have gone by, with $t = 0$ being the time that the first test began, then we guarantee

$$E[\mathsf{FDR}_T] \leq q$$

with q a pre-determined threshold, and T a stopping time, or rule describing a decision to stop and look at the matrix of tests. One example of such a decision T is "report results when 5 variations x goal pairs all have FDR below 10%." As before, regardless of when you look at the matrix, if you call the tests with $\mathsf{FDR} \leq q$, then you are guaranteed to control false discovery rate at level q.

4 Technical appendix: Sequential testing with likelihood ratios

Here is the full derivation of the threshold on θ_n given in (2). To do this, we change perspective from θ_n to the likelihood function. Let $X_1, X_2, \ldots$ denote the sequence of observations for visitors in the baseline group (where a one means the visitor converted, and zero means they did not); and let $Y_1, Y_2, \ldots$ denote the sequence of observations for visitors in the variation group. Given any candidate θ for the true conversion rate difference, the likelihood of θ based on the first n observations is defined as:

$$L_n(\theta; \hat{\theta}_n) = \mathtt{P}_\theta(X_1, ..., X_{n/2}, Y_1, ..., Y_{n/2})$$

where $\mathtt{P}_\theta$ denotes probabilities when the true difference is θ, and $\hat{\theta}_n$ is a sufficient statistic encoding all the information needed from the first n observations to calculate L_n. Until the end of this section, we drop the parameter $\hat{\theta}_n$ for ease of explanation, because it is not needed fo the arguments here. It becomes important again in section 5, where we combine our results.

The likelihood function is a fundamental quantity that is of central importance across statistics. In any statistical model, it formalizes the plausibility of a candidate parameter value θ, given the data observed.

Recall that our overall objective comprises two parts: to adaptively learn the effect size we should anticipate, and then to stop as soon as possible to detect that effect when it holds. We begin by isolating the latter goal. Suppose that we knew that the true conversion rate difference was either zero or some given non-zero alternative $\tilde{\theta}$. Then the likelihood ratio of $\tilde{\theta}$ against zero is defined as

$$\Lambda_n(\tilde{\theta}) = \frac{L_n(\tilde{\theta})}{L_n(0)}$$

This represents the relative plausibility of the alternative.

At the start of the experiment $L_0(\tilde{\theta}) = L_0(0) = 1$ so this ratio is 1. Over the course of the experiment, the value of $\Lambda_n(\tilde{\theta})$ will fluctuate as θ_n does. In fact $\Lambda_n(\tilde{\theta})$ depends on the data only through the value of θ_n (that is the sense in which θ_n is "sufficient") so placing a threshold on θ_n is equivalent to placing a threshold on Λ_n. But unlike θ_n, Λ_n does not converge to zero when the variation and the baseline are the same. Rather, an important property of the likelihood ratio is that, if the true difference is zero, then $\Lambda_n(\tilde{\theta})$ is a martingale. This means that, conditional on the value Λ_n we observe after n visitors, at a later stage $m > n$ the expected value of Λ_m is still Λ_n; there is no average drift upwards or downwards.

Without drift, we can get to our goal just by placing a constant threshold on Λ_n. For any positive martingale started at 1, the probability of hitting a constant threshold $b > 1$ is less than $1/b$. So the rule: *stop when* $\Lambda_n(\tilde{\theta}) \geq 1/\alpha$ gives always valid Type I error control at level α.

The proof of this fact (which relies on the Optional Stopping Theorem) is roughly as follows. Let T_b be the least n at which $\Lambda_n \geq b$ is achieved, with $T_b = \infty$ if the threshold is never crossed. Fix some maximum number of visitors t and let $T_b \wedge t$ be the lesser of T_b and t. After $T_b \wedge t$ visitors arrive the expected value of Λ_n is still its initial value 1. But, by construction,

$$\Lambda_{T_b \wedge t} \geq \begin{cases} b & T_b \leq t \\ 0 & T_b > t \end{cases}$$

so

$$1 = \mathtt{E}_0 \, \Lambda_{T_b \wedge t} \geq b \, \mathtt{P}_0(T_b \leq t)$$

Taking the maximum visitors $t \to \infty$ bounds the probability of ever crossing the threshold at

$$\mathtt{P}_0(T_b < \infty) \leq 1/b$$

We want to extend this approach so that the true difference can take any value. This formula is at the heart of the formulation of a sequential test as we now can control the probability that our likelihood ratio EVER crosses the threshold b. It is important to note however, that a constant threshold on a single $\Lambda_n(\tilde{\theta})$ is not good enough, because when we map back to θ_n the resulting threshold does not converge to zero. For example, while we are guaranteed to detect an effect of size $\tilde{\theta}$ if it exists, we are not guaranteed to detect smaller effects like $\tilde{\theta}/2$ or $\tilde{\theta}/5$. The target is to consider the likelihood ratio of every potential effect θ against zero simultaneously, and hone in on the appropriate candidate. This motivates a Bayesian approach, where we place a prior, $\pi(\theta)$, on the true effect.

Then we can monitor the average likelihood ratio

$$\Lambda_n(\hat{\theta}_n) = \int \Lambda_n(\theta) \, \pi(\theta) \, d\theta$$

which represents the relative plausibility of the aggregate alternative that considers every θ, weighted according to $\pi(\theta)$. Again we highlight the fact that our final statistic is a function of the data $\hat{\theta}_n$. In Bayesian terminology, Λ_n is called the *Bayes factor* in favor of the alternative hypothesis. Any mixture of martingales is itself a martingale (if each statistic tends to stay where it is, then so does the average). So, as before, we can get an always valid threshold on θ_n by placing a constant threshold of $1/\alpha$ on Λ_n.

We choose a normal prior

$$\pi(\theta) \sim N(0, \tau)$$

both for computational convenience and because central limting arguments make the distribution of effect sizes effectively normal for experiments run on any even moderately popular A/B testing platform.

Using this prior and integrating, we obtain the boundary in (2).

Given thresholds, it is easy to convert these tests into p-values. A p-value is the most conservative significance level at which the test is conclusive, so we take the p-value after n visitors to be the least α such that the threshold has already been crossed. As the experiment progresses, more thresholds are crossed and the p-value decreases, approaching some asymptotic value at $n = \infty$. In terms of p-values, the statement that *the probability that the α-level test is ever conclusive is* α is equivalent to saying that the asymptotic p-values are uniformly distributed. So, at $n = \infty$ the p-values are as aggressive as they can be, while still controlling Type I error. However during the experiment, the p-value is an overestimate of this asymptotic p-value so we are slightly more conservative: this can be considered the mathematical penalty we must pay to be free to stop tests as early as possible.

5 Reference Appendix

Traditional, fixed-horizon statistics

- $\hat{p} = p(\hat{\theta}_n, \theta')$, traditional p-value for evidence against the null hypothesis, $H_0 : \theta = \theta'$
- $C(\hat{\theta}_n, \alpha) = \{\theta \mid p(\hat{\theta}_n, \theta) \geq \alpha\}$, traditional confidence interval with $1 - \alpha$ coverage level

Stats Engine statistics

- $\Lambda(\hat{\theta}_n)$, average likelihood ratio; inverse of new p*-value
- $q^*(\hat{\theta}_n) = BHQ^*(\frac{1}{\Lambda(\hat{\theta}_n)})$, FDR-adjusted p*-value
- $C^*(\hat{\theta_n}, \alpha) = \{\theta \mid \Lambda(\hat{\theta}_n) < 1/\alpha\}$, new confidence interval with $1 - \alpha$ coverage
- $Q^*(\hat{\theta}_n, \alpha) = C^*(\hat{\theta}_n, FCR(\alpha))$, FCR-adjusted coverage interval

Shown on Optimizely dashboard

- $q^*(\hat{\theta}_n)$, as "Statistical Significance"
- $Q^*(\hat{\theta}_n, \alpha)$, as numerical and graphical coverage interval
- α, threshold for declaring winners ($\hat{\theta}_n > 0$) and losers ($\hat{\theta}_n < 0$), by $q(\hat{\theta}_n) < \alpha$, set in account level settings

References

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

[2] Y. Benjamini and D. Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

[3] J. M. Hoenig and D. M. Heisey. The abuse of power. *The American Statistician*, 55(1), 2001.

[4] T. Lai. Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, pages 303–408, 2001.

[5] H. Robbins. Statistical methods related to the law of the iterated logarithm. *Ann. Math. Statist.*, pages 1397–1409, 1970.

[6] D. Siegmund. *Sequential Analysis: Tests and Confidence Intervals.* Springer, 1985.

[7] A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945.