False Positives in A/B Tests

Ron Kohavi
Kohavi
Los Altos, CA
ronnyk@live.com

Nanyu Chen
Expedia Group
San Francisco, CA
nc361sirg@gmail.com

ABSTRACT

A/B tests, or online controlled experiments, are used heavily in the software industry to evaluate implementations of ideas, as the paradigm is the gold standard in science for establishing causality: the changes introduced in the treatment *caused* the changes to the metrics of interest with high probability. What distinguishes software experiments, or A/B tests, from experiments in many other domains is the scale (e.g., over 100 experiment treatments may launch on a given workday in large companies) and the effect sizes that matter to the business are small (e.g., a 3% improvement to conversion rate from a single experiment is a cause for celebration). The humbling reality is that most experiments fail to improve key metrics, and success rates of only about 10-20% are most common. With low success rates, the industry standard alpha threshold of 0.05 implies a high probability of false positives. We begin with motivation about why false positives are expensive in many software domains. We then offer several approaches to estimate the true success rate of experiments, given the observed "win" rate (statistically significant improvements), and show examples from Expedia and Optimizely. We offer a modified procedure for experimentation, based in sequential group testing, that selectively extends experiments to reduce false positives, increase power, at a small increase to runtime. We conclude with a discussion of the difference between ideas and experiments in practice, terms that are often incorrectly used interchangeably.

CCS CONCEPTS

General and Reference → Cross-computing tools and techniques → Experimentation; Mathematics of computing → Probability and statistics → Probabilistic inference problems → Hypothesis testing and confidence interval computation

KEYWORDS

A/B Testing, Controlled experiments.

ACM Reference format:

Ron Kohavi, Nanyu Chen. False Positives in A/B Tests. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25-29, 2024, Barcelona, Spain. ACM, New York, NY, USA.

1 INTRODUCTION

A/B tests, or online controlled experiments [1; 2; 3; 4; 5; 6; 7; 8; 9], are used heavily in the software industry to evaluate implementations of ideas, as the paradigm is the gold standard in science for establishing causality: the changes introduced in the treatment *caused* the changes to the metrics of interest with high probability.

What distinguishes software experiments, or A/B tests, from experiments in many other domains is the (e.g., over 100 experiment treatments may launch on a given workday in large companies) [1; 10; 11; 12; 13] and the effect sizes that matter to the business are small (e.g., a 3% improvement to conversion rate from a single experiment is a cause for celebration).

The humbling reality is that most experiments fail to improve key metrics, and success rates of only about 10-20% are most common [14; 1]. With low success rates, the industry standard alpha of 0.05, which is the p-value threshold for statistical significance, implies a high probability of false positives [14; 15; 16; 17; 11; 18].

A false positive is defined as a statistically significant result (H_0 is rejected) when H_0 should not have been rejected; that is, given our sample size, the true treatment effect is not inconsistent with H_0. We are not assuming that the treatment effect is exactly zero, as outside limited scenarios, it is unlikely that any change will have exactly no effect [19], but we are asking if our observations are consistent with zero effect [15].

The paper is structured as follows: we begin with motivation about why false positives are more expensive than false negatives in many software domains. In Section 3, we review the estimation of false positive risk and the need to estimate the true success rate from the observed "win" rate. In Section 4, we offer several approaches to estimate the true success rate of experiments and show examples from Expedia and Optimizely. In Section 5, we look at the False Positive Risk for a specific p-value or narrow range. In Section 6, we offer a modified procedure for experimentation, based on sequential group testing, that selectively extends experiments to reduce false positives, increase power, at a small increase to runtime. In Section 7, we discuss the difference between ideas and experiments in practice, two terms that are often incorrectly used interchangeably. We conclude with a summary.

2 THE COSTS OF FALSE POSITIVES AND FALSE NEGATIVES

We begin with an example of a highly expensive false positive in Psychology. It is now well agreed that many of the most cited psychology findings failed to replicate [20]. One example is that of Ego-Depletion, a psychological theory based on a 1998 paper that claims we have a limited pool of mental resources that we use up, after which we lose self-control [21]. Schimmack claimed that Ego-Depletion suffers from low replicability [22]. A large multisite preregistered test with 36 laboratories and 3,531 participants found a nonsignificant effect with d=0.06, an effect about 10 times smaller than believed [23]. Given these results, the initial result was a false positive and the effect, if it exists, is tiny. The initial study held for over 15 years with widespread confidence in the robustness of the effect, including a meta-analysis of 198 independent tests in 2010. What a waste of resources! In hindsight, this showed how much bias there is in accepted publications, where non-significant results are often not submitted or rejected (the file drawer problem [24]), and statistically significant results are published.

What about in A/B tests, or online experiments? While there are certainly cases where experiments are run for short-term decisions (e.g., headline optimizations), for most experiments the real impact of false positive results is on the roadmap, or the backlog—steering the ship into the wrong direction because of some amazing discovery that is wrong: a false positive.

False negatives also have a cost, of course. We may have a good idea, which will go unnoticed. There are two reasons why we believe these costs are lower: (1) for an idea to be a false negative, it is likely near our MDE (minimum detectable effect) that we use to determine the sample size, as the probability of failing to detect breakthrough ideas, with a large effects, is very small; (2) organizations typically run a few experiments with small variations before giving up on an idea, so the probability of all these variations being false negatives diminishes exponentially fast (e.g., with the industry standard 20% type-II error, the probability of 5 variations failing to be statistically significant is 0.2^5=0.03%).

When we choose parameters like alpha, the threshold for rejecting the null hypothesis, we are trading off false positives with false negatives. While every organization can determine the costs, in Section 4.5 on choosing alpha, we share both a 1:1 cost and a 3:1 cost, where false positives are three times more expensive.

3 ESTIMATING THE FALSE POSITIVE RISK

P-values are commonly misinterpreted as the probability of making a mistake when choosing the Treatment over Control when the observed metric of interest is statistically significantly different [25; 26; 27]. Multiple examples of this misinterpretation by A/B vendors, book authors, and in courts were provided by Kohavi, Deng, and Vermeer [14].

What is the p-value then? The p-value is the probability of obtaining a result equal to or more extreme than what was observed, assuming that *all* the modeling assumptions, including the null hypothesis, H_0, are true [26]. Conditioning[1] on the null hypothesis is critical and most often misunderstood. In probabilistic terms, we have

$$\text{p-value} = P(\Delta \textit{ observed or more extreme} | H_0 \textit{ is true}) .$$

What we are looking for most of the time is the opposite conditional probability:

$$P(H_0 \textit{ is true } | \Delta \textit{ observed})$$

Using Bayes Rule, we can estimate the False Positive Risk (FPR), which is the probability that the statistically significant result is a false positive, or the probability that H_0 is true (no real effect) when the test was statistically significant [15]. Note that FPR is sometimes named FDR, or False Discovery Rate [28; 29], but given the confusion with FDR from multiple hypothesis testing, we use the term recommended by Colquhoun [15].

We use the following terminology [14]:

1. **SS** is a statistically significant result.
2. $\boldsymbol{\alpha}$ is the threshold used to determine statistical significance (SS), commonly 0.05 for a two-tailed t-test.
3. $\boldsymbol{\beta}$ is the type-II error (usually 0.2 for 80% power)
4. $\boldsymbol{\pi}$ is the prior probability of the null hypothesis, that is $P(H_0)$

we can apply Bayes Rules for the following:

$$FPR = P(H_0|SS) = \frac{\alpha * \pi}{\alpha * \pi + (1-\beta) * (1-\pi)} .$$

An alternative derivation of FPR, resulting in the same formula, was made in the Supplement to Equation 2 and Figure 2 in Benjamin et. al. [30].

The key parameter required for the above is $\boldsymbol{\pi}$, or $P(H_0)$. Kohavi, Deng, and Vermeer [14] provided a table with seven success rate estimates (1- $\boldsymbol{\pi}$) that were reported in the software industry, which ranged from 8% to 33% with a median and mode of 10%. Plugging these into the above formula results in an FPR of 22% for the median and mode success rate of 10%, industry standard alpha of 0.05, and 80% power. This is a much higher rate than people intuitively think of when they hear statistically significant improvement. For companies that use α=0.10 as their threshold

[1] Some authors prefer to use the semicolon notation; see discussion at: https://statmodeling.stat.columbia.edu/2013/03/12/misunderstanding-the-p-value/#comment-143481

for statistical significance, or equivalently use α=0.05 with a one-tailed test for the improvement tail (e.g. Optimizely [31], Analytics Toolkit [32], Booking.com [33], Expedia), the FPR for 10% success rate is a 36%. Over one third of the statistically significant results showing improvement, which we want to celebrate, are likely to be false positives!

To provide intuition about why the FPR is so high when the success rate is low, we will use the data reported by Optimizely [34] of 12% win rate across 127,000 experiments. As we will show later in the paper in Section 4.4, the estimated true success rate is 9.3%, in line with the 10% median and mode of Table 2 in Kohavi, Deng, and Vermeer [14].

Looking at Figure 1, the dot-pattern (also green if viewed in color) in the first row represents the 9.3% success rate, that is, true effects that should be statistically significant given our sample size providing 80% power. Of these, 80% will be identified as statistically significant, so 80%*9.3%=7.4% are denoted by a plus in the first row.

Of the remaining 90.7% null effects, 5% will be statistically significant and positive, so 4.5% of the A/B tests will show a statistically significant result: a false positive. These are denoted by a plus in the second row.

Of the ~12% wins (7.4%+4.5% depicted by pluses), 4.5% are false positives, so 4.5%/(4.5%+7.4%)=37.8%. This surprisingly high false positive is often referred to as the base rate fallacy [35].

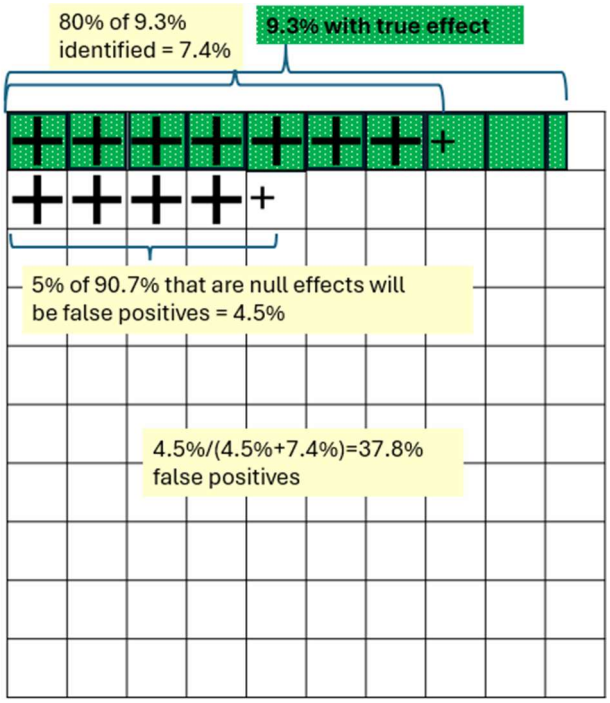

Figure 1: Diagram showing how success rate of 9.3% implies a false positive risk of 37.8%.

4 ESTIMATING THE SUCCESS RATE

When we apply the Null Hypothesis Significant Testing (NHST) methodology with a null hypothesis that the effect size is zero, we are not claiming that the effect is exactly zero; rather, we build a statistical model that assumes it is zero, and we ask whether the results are inconsistent with that effect size. This is sometimes referred to as the champion/challenger model, where the current champion (Control) is maintained until we can show that a challenger (Treatment) is very likely different[2]. If the results are inconsistent with an effect size of zero, then we reject the null hypothesis and estimate the effect size [15]. For a treatment to be successful, it must meet two conditions:

a. The effect size must be large enough to reject the null hypothesis at an alpha level of choice (e.g., 0.05).
b. The direction of the effect must be such that it improves the metric of interest. For example, for conversion rate or revenue, a positive delta is desired, whereas for performance (e.g., time to generate a page, or some elements of the page), a negative delta is desired.

Here we must differentiate between an observed success, which we call a "win," and (a true) success, which implies that the underlying true effect is large enough that we would reject the null hypothesis at the given sample size and chosen alpha level.

In a single experiment, we may observe a value that is lower or higher than the true treatment effect, which could result in a false negative or false positive statistically significant result. It is worth noting that with the standard design of 80% power, the expected p-value is 0.005, or equivalently Z=2.8, for a true effect that matches the MDE. Because we declare a result statistically significant below p-value of 0.05, or Z-score > 1.96, we will declare an observed result greater than 1.96/2.8=70% of the MDE as statistically significant.

A **false positive statistically significant result** is one where we have a "win" because we computed a p-value less than alpha and can thus declare statistical significance, but the true effect size is consistent with the null hypothesis, given our sample size. For example, let's assume that we observe a treatment effect, delta, and the p-value computation comes out at 0.02 (statistically significant). If the true treatment effect is smaller, say 0.6*delta, and the p-value for that effect is greater than 0.05, then we have a false positive result because true effect size is consistent with the null hypothesis, given our sample size. Experiments with low power that are statistically significant, tend to exaggerate the effect size [36]. For example, an experiment with 20% power for a given MDE will not be statistically significant 80% of the time;

[2] We are assuming the goal is to improve the Control. There are scenarios not discussed here, where success is non-inferiority (e.g., legal requirement).

in 20% of the cases, it will be statistically significant, but the treatment effect will be exaggerated by a factor of 2.3 or 130% [37], often referred to as the winner's curse [38; 39; 40].

A **false negative result** arises when the p-value of the true effect size is less than alpha, but the p-value computed for the observed treatment effect is greater than alpha. If our sample size is chosen to provide 80% power for a given MDE (minimum detectable effect) and level of alpha, and if the true effect matches the MDE, then 20% of the time we will have a false negative, or type-II error: we will not have a "win" even though we should have. Given an observed win rate, how can we estimate the true success rate? We now review three approaches.

4.1 The Naïve Approach

In the Naïve Approach, we simply estimate the success rate as the win rate, or the rate of statistically significant improvements. This approach ignores false positives at a rate of FPR (note, NOT type-I error rate, which assumes the null hypothesis is true), but also ignores the type-II error rate (e.g., missing 20% of the true successes). These two factors are in opposite directions, so their relative magnitude matters. This was the approached used to estimate the success rate at Microsoft [41] and Bing [18].

4.2 Replicated Experiments

In this approach, experiments with borderline p-values (e.g., 0.01-0.10) are replicated and the two p-values are combined using Fisher's method or Stouffer's method [42 p. Fisher's method] thus increasing the statistical power significantly (reducing type-II error), while reducing false positives. The threshold for the combined p-value could be set at 0.01, for example. If the combined result is still borderline, a third experiment may be required, although that seems quite rare in practice. Table 1 shows examples of two p-values and the combined result using Stouffer's method, resulting in a combined p-value less than 0.01 [43]:

Table 1: Meta analysis: combining p-values from two experiments.

P-value 1	P-value 2	Combined P-value
0.06	0.07	0.009
0.05	0.09	0.009
0.04	0.10	0.009
0.03	0.14	0.0093
0.02	0.18	0.0095
0.01	0.28	0.0097

At Airbnb, we were surprised to see some replication runs that failed, and that led to the important realization that the true success rate was lower than we had assumed from our winning (statistically significant improvement) experiments. We implemented the above approach to approximate the success rate and estimated it at 8% [14]. While this approach significantly reduces false positives by re-testing, it doesn't account for false negatives, so the success rate estimated is conservative.

4.3 Conditional Probabilities

We can decompose P(SS), the probability of a statistically significant result, by conditioning on the null hypothesis, and isolate the success rate as follows:

$$P(SS) = P(SS|H_0) * P(H_0) + P(SS|\neg H_0) * P(\neg H_0)$$
$$P(SS) = \alpha * \pi + (1-\beta) * (1-\pi)$$
$$P(SS) = \alpha * \pi + 1 - \beta - \pi + \beta\pi$$
$$P(SS) = \pi(\alpha + \beta - 1) + 1 - \beta$$
$$\pi\,(1 - \beta - \alpha) = 1 - \beta - P(SS)$$
$$\pi = \frac{1 - \beta - P(SS)}{1 - \beta - \alpha}$$
$$\pi = \frac{\text{Power} - P(SS)}{\text{Power} - \alpha}$$

Table 2 shows how the observed statistically significant rate at different alpha levels translates into an estimate of the true success rate, assuming power is 80%. The spreadsheet is available at https://bit.ly/FalsePositivesInABTestsCalc (success rate estimation tab).

Table 2: Success rate estimation using conditional probabilities.

Win-rate (observed stat-sig improve-ment)	Alpha (two-tailed)	True success rate	False positive risk (FPR)	False positiv es (of all exper-iments)	False negatives (of all exper-iments; Type-II error)
30%	0.2	28.6%	23.8%	7.1%	5.7%
30%	0.1	33.3%	11.1%	3.3%	6.7%
30%	0.05	35.5%	5.4%	1.6%	7.1%
30%	0.01	37.1%	1.0%	0.3%	7.4%
20%	0.2	14.3%	42.9%	8.6%	2.9%
20%	0.1	20.0%	20.0%	4.0%	4.0%
20%	0.05	22.6%	9.7%	1.9%	4.5%
20%	0.01	24.5%	1.9%	0.4%	4.9%
15%	0.2	7.1%	61.9%	9.3%	1.4%
15%	0.1	13.3%	28.9%	4.3%	2.7%
15%	0.05	16.1%	14.0%	2.1%	3.2%
15%	0.01	18.2%	2.7%	0.4%	3.6%
12%	0.2	2.9%	81.0%	9.7%	0.6%
12%	0.1	9.3%	37.8%	4.5%	1.9%
12%	0.05	12.3%	18.3%	2.2%	2.5%
12%	0.01	14.5%	3.6%	0.4%	2.9%
10%	0.2	0.0%	100.0%	10.0%	0.0%
10%	0.1	6.7%	46.7%	4.7%	1.3%

10%	0.05	9.7%	22.6%	2.3%	1.9%
10%	0.01	11.9%	4.4%	0.4%	2.4%
8%	0.1	4.0%	60.0%	4.8%	0.8%
8%	0.05	7.1%	29.0%	2.3%	1.4%
8%	0.01	9.4%	5.7%	0.5%	1.9%

At the industry standard alpha of 0.05, the Naïve Approach happens to be a reasonable approximation for 10-15% win rates. For example, observing 15% statistically significant improvements (top highlighted row in the table) is reasonably close to the 16.1% true success rate. This is because in 83.9% of the time (100%-16.1% success rate), when the null hypothesis is true, our type-I error rate of 2.5% will generate 2.1% (83.9%*0.025) false positive statistically significant positive results, and in the remaining 16.1% of the time, our 80% power will generate 12.9% (16.1%*0.8) statistically significant result, and thus 2.1%+12.9%=15% will be "wins," that is, statistically significant improvements will be observed. The FPR column is a good sanity check: for our 15% statistically significant positive results, 14% are false positives, so 15%*14%=2.1%, exactly matching the computation above.

At higher and lower alphas, the Naïve approach is poor. For example, let us review the bottom, highlighted row, where the observed stat-sig rate is 10% and alpha is 0.2. The true success rate is estimated at 0%. The reason is that in the scenario where the null hypothesis is always true, that is, the treatment effect is consistent with zero, then with alpha=0.2, 20% of the time we will make a type-I error and declare a statistically significant result. Half of these will be in the improvement direction, so all 10% observed statistically significant improvements are false positives. The FPR is indeed 100%: every statistically significant result is a false positive.

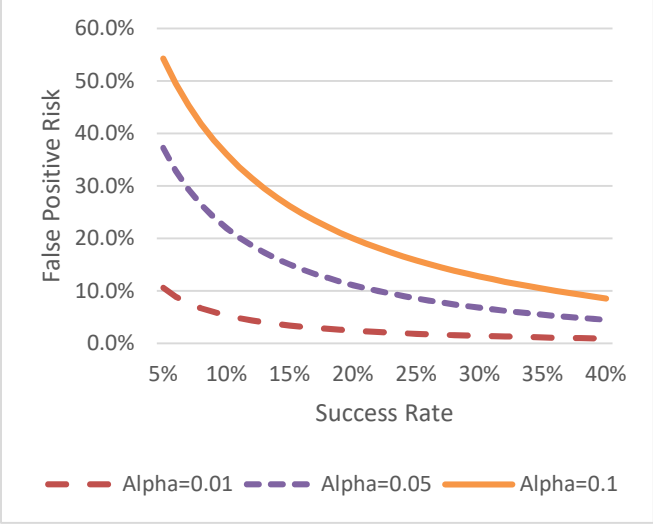

Figure 2: False Positive Risk for different levels of Alpha (two-tailed), 80% power.

Figure 2 shows the relationship between FPR and success rate for different levels of alpha. At low success rates, near the median of 10% for reported success rates in Kohavi, Deng, and Vermeer [14], higher alpha result in very high false positive risks, even for 80% power.

4.4 True Success Rate and FPR for Optimizely's Customers

Optimizely, a vendor of A/B testing software, recently published a nice report with lessons learned from 127,000 experiments [34]. They reported that the average win rate for the primary metric of these experiments was 12%, when Optimizely's default for alpha is 0.10 [31]. From Table 2, the middle-highlighted row, we can estimate the following:

a. The true success rate is 9.3%.
b. The FPR, or False Positive Risk, is 37.8%.

The 9.3% is close to multiple reports of 10% success rates [14], but because of the high default alpha, which is double the industry standard of 0.05, the probability that a statistically significant result reported by Optimizely is a false positive is over a third! Optimizely uses the term "90% confidence" for alpha=0.10, and it is very likely that their users believe their probability of a false positive is 10%, not 37.8%.

4.5 True Success Rate and FPR at Expedia

We evaluated Expedia's win rate over thousands of trustworthy experiments. Expedia historically used two-sided tests with alpha=0.10, and for that threshold, the win rate was 15.6%. Like the above computation for Optimizely, we can estimate the following:

a. The true success rate is 14.1%.
b. The FPR, or False Positive Risk, is 27.5%.

While not as high as Optimizely, the FPR is still high, and Expedia is lowering alpha.

4.6 Choosing Alpha

Given a set of experiments with data about them, including the number of users (more generally, units) and the p-value for each, how should an organization set the appropriate alpha level?

Bartoš and Schimmack [29] suggested to use the curve of Z-scores to arrive at the replication probability of a hypothesis. The key observation is that if experiments are powered at 80%, the mean Z-score should be about 2.8, which corresponds to a p-value of 0.005. If most results are borderline statistically significant with p-values around 0.01-0.05, for example, then there is a strong publication bias and results suffer from the file drawer problem, where many results that are not statistically significant are not submitted or not accepted for publication [24; 44]. The file drawer problem should not exist in online experimentation platforms, which typically track all results for the organization. This is likely

the reason that reported success rates are around 10-20%, relative to 85% in medical journals and 95% in Psychology journals [44].

Indeed, an experimental summary from Bing [45] presented in Figure 3 shows the experiments' treatment effects. Not only is the mass concentrated around zero, but there are many negative effects, even though most experiments are designed to evaluate new features to improve the experience.

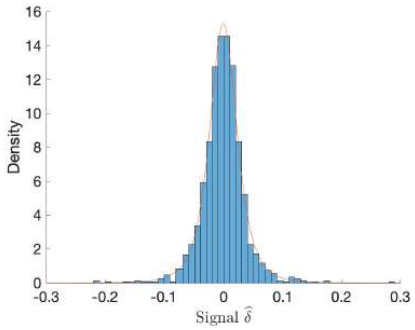

Figure 3: Histogram of success rate and estimated density [45].

Similarly, a summary of 1,001 A/B tests that have used Analytics-toolkit.com [46] is presented in Figure 4. The median lift is at 0.1%, with a large mass concentrated around zero.

Table 3 shows the true success rate, alpha, power (keeping sample size fixed and powered at 80% for alpha=0.05), false-positives, and false-negatives that are implied. The last two columns show equal-weighted cost of 1:1, perhaps reasonable for short-term decisions like headlines and 3:1 cost, which may be more appropriate for ideas evaluated towards learning. For each group of true success rate, the alpha minimizing the cost is highlighted. For these costs, it is clear that many organizations should lower alpha to 0.01-0.05. We provide the spreadsheet at https://bit.ly/FalsePositivesInABTestsCalc (success rate estimation tab, bottom) so that you can plug your own costs. It's also possible to optimize with fixed power at various alphas by taking into account the impact on experiment velocity, along with implied false positive and negatives.

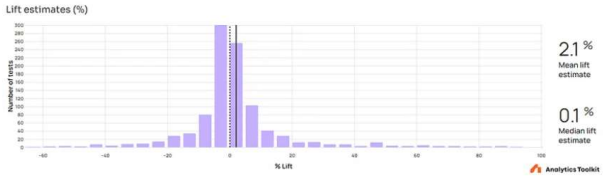

Figure 4: Distribution of lift estimates for 1,001 A/B Tests [46].

Our advice to lower alpha in many scenarios is aligned with others who have also recommended lowering alpha rather than increasing power [28].

Table 3: alpha choices and their cost for two different cost functions: 1:1 assigns 1 to false positives and 1 to false negatives 3:1 assigns 3 to false positives and 1 to false negatives

True success rate	Alpha (two-tailed)	Power	False positives	False negatives (Type-II error)	1:1 cost	3:1 cost
30.0%	0.2	94%	7.0%	1.9%	8.9%	22.9%
30.0%	0.1	88%	3.5%	3.7%	7.2%	14.2%
30.0%	0.05	80%	1.8%	6.0%	7.8%	11.3%
30.0%	0.01	59%	0.4%	12.3%	12.7 %	13.4%
20.0%	0.2	94%	8.0%	1.3%	9.3%	25.3%
20.0%	0.1	88%	4.0%	2.5%	6.5%	14.5%
20.0%	0.05	80%	2.0%	4.0%	6.0%	10.0%
20.0%	0.01	59%	0.4%	8.2%	8.6%	9.4%
15.0%	0.2	94%	8.5%	1.0%	9.5%	26.5%
15.0%	0.1	88%	4.3%	1.9%	6.1%	14.6%
15.0%	0.05	80%	2.1%	3.0%	5.1%	9.4%
15.0%	0.01	59%	0.4%	6.2%	6.6%	7.4%
12.0%	0.2	94%	8.8%	0.8%	9.6%	27.2%
12.0%	0.1	88%	4.4%	1.5%	5.9%	14.7%
12.0%	0.05	80%	2.2%	2.4%	4.6%	9.0%
12.0%	0.01	59%	0.4%	4.9%	5.4%	6.3%
10.0%	0.2	94%	9.0%	0.6%	9.6%	27.6%
10.0%	0.1	88%	4.5%	1.2%	5.7%	14.7%
10.0%	0.05	80%	2.3%	2.0%	4.3%	8.8%
10.0%	0.01	59%	0.5%	4.1%	4.6%	5.5%
8.0%	0.1	88%	4.6%	1.0%	5.6%	14.8%
8.0%	0.05	80%	2.3%	1.6%	3.9%	8.5%
8.0%	0.01	59%	0.5%	3.3%	3.8%	4.7%

5 FALSE POSITIVE RISK for P-VALUE

We previously looked at the FPR, or False Positive Risk, for all experiment results less than alpha. Experiments with lower p-value are likely to have a lower FPR than experiments with a higher p-value closer to alpha. Colquhoun [15] refers to these two cases as less-than-alpha, and equal-to p-value.

5.1 Motivating Example

Optimizely reported a win rate of 12% in their analysis of 127,000 experiments [34]. Assuming these experiments were properly run with 80% power (it would be worse if these were under-powered) and using their default alpha of 0.10 [31], the overall FPR, or False

Positive Risk, is 37.8% as shown in Section 4.3. This estimate is computed for all experiment results less than alpha.

Assume an e-commerce site that uses Optimizely, has 5% conversion rate, and powers up an experiment to detect an MDE of relative 5%, which is a scaled version of Colquhoun example in his paper [15]. If we look at a specific experiment that has a p-value right at alpha, so p-value = 0.10 in the case of Optimizely, we can estimate a more specific FPR. Colquhoun's simulations for our case show an FPR for the less-than-alpha case of 37.6%, very close to the Bayes Rules estimate of 37.8% mentioned above. As you might expect, all results less-than-alpha, have a lower FPR than a p-value equal to alpha, as lower p-values indicate more evidence against the null hypothesis. If the p-value for a specific experiment is 0.10 (90% confidence in Optimizely's terminology [31]), Colquhoun's simulation show that the probability of a false positive is 80.7% (see Appendix A for screen shot from the simulator). In this case, a statistically significant result, which Optimizely will say has 90% confidence, will be a false positive over 80% of the time!

5.2 Estimating FPR for P-value

We would like to estimate the FPR for p-value in the range [p1,p2]. We will use the notation P12 to indicate the probability Pr($p1 \leq$ p-value $\leq p2$).

$$FPR(p1,p2) = P(H0|\,P12) = (P12\,|\,H0)\cdot P(H0)\,/\,P12$$

For simple (non-composite) null hypotheses, such as delta=0, the p-value distribution is uniform [47], so we have

$$= (p2-p1)\cdot\pi\,/\,P12$$

The denominator can be estimated by the proportion of the number of experiments between p1 and p2 to the overall count of experiments. Practically, the range must be large enough so that there are enough experiments for the estimate to have low variance.

We use Expedia's success rate derived in section 4.5, along with the perturbed empirical distribution of p-value of tests, to demonstrate the FPR at different p-value ranges typical seen in large-scale organizations running thousands of tests. Table 4 shows the FPR for three ranges. The spreadsheet at https://bit.ly/FalsePositivesInABTestsCalc "FPR for p-value" tab has the computations for the table.

Table 4: FPR for P-value ranges

p-value range	Numerator (False positives%)	Percent of experiments in this range (perturbed)	False Positive Risk
[0, 0.01]	0.4%	7.2%	6.0%
[0.01, 0.05]	1.7%	5.2%	33.0%
[0.05, 0.1]	2.1%	3.2%	67.1%

It is clear that experiments with p-value between 0.05 to 0.10 have a surprisingly large probability of being false positives, as the FPR is 67%.

6. Replicating or Extending the Experiment

In Section 4.2, we suggested replicating experiments to help assess the true success rate. In this section we analyze the cost of using this procedure or a similar alternative of extending the experiment, not for estimating the success rate, but rather on an ongoing basis.

In the rest of this section, we will assume that the common industry standard of alpha=0.05 is being used with a true success rate of 10%, which is the median and mode from Kohavi, Deng, and Vermeer [14]. For this case, the FPR for less-than alpha is 22% and the Colquhoun's simulation for p-value equal case is 64%. For Expedia, the p-value equal case is 37% using 0.05 ∓ 0.1 as the range (described in Section 5.2).

Given the high FPR of experiments with a p-value close to the alpha threshold, we can decide to validate the results by doing a replication run, or by extending the experiment if the p-value is close to alpha, say < 0.05. One option is to extend the experiment by the same duration it originally ran (e.g., two weeks). For simplicity, we will assume that this will double the number of users, although in practice user growth is sub-linear due to repeat users (the calculations can be similarly done if the first two weeks represents, say, 60% of the total users vs. 40% in the latter two weeks, or the extension can be for, say, three weeks).

Another option is to shuffle the users and start a replication run. Unlike extending the experiment, there is no issue with the count of repeat users, and the number of users over the second set of two weeks will be similar to the first two weeks, as users are re-randomized. Deng, Li, and Guo [48] show weak dependence from such a replication run, which implies that a combined analysis can be done. This approach is "cleaner" from a statistical perspective, but if the experiment is clearly visible to users, such as an obvious UI (User Interface) change, then the first approach is preferable so that users aren't contaminated by seeing both Control and Treatment. We can do a meta-analysis of the two experiments using Fisher's method or Stouffer's method [42 p. Fisher's Method], but a better approach is to view this as a group-sequential test.

Using alpha spending from group sequential tests [49; 50], we can design a protocol with interim analysis, where the type-I error rate is still restricted to some alpha, such as 0.05. For example, using the ldbounds package in R, here are three options:

5. Selecting the Pocock approach [51], an analysis at 50% of users with alpha1=0.03 and an analysis at 100% of users with alpha=0.03 keeps the overall alpha at 0.05.
6. Selecting O'Brien-Fleming [52], is conservative at the midpoint, declaring success only if the p-value is less than alpha1=0.005, yet still be able to declare statistical

significance at the end if the p-value is less than alpha2=0.048.

7. Goldilocks (in between the above): setting alpha1 to 0.01, and thus alpha2 should be <= 0.046 to control for overall type-I error of 5%.

Note that due to repeat users, a more sophisticated analysis can be done than the above package by providing the covariance matrix, but the above is likely to be conservative.

In terms of power for the final (e.g., four-week) experiment, not doing any interim analysis is the most powerful. O'Brien-Fleming loses a bit of power with alpha2=0.048, but the probability of being able to stop after two weeks is low. The Goldilocks approach seems like a reasonable tradeoff where more experiments can be terminated at two weeks without losing much power.

If the initial experiment (e.g., two weeks) was powered at 80% power for some MDE, then doubling the users reduces the MDE by a factor of $\sqrt{2} = 1.41$, or, equivalently, raising the power of the original MDE from 80% to about 98% ($1 - \beta = \Phi(\delta/\text{SE} - Z_{1-\alpha/2})$ [14], where $\delta = 2.8SE1$ and $SE = SE1/\sqrt{2}$, so $1 - \beta = \Phi(2.8\sqrt{2} - 1.995) = 98\%$, where 1.995 comes from the Goldilocks 0.046 threshold). Given 80% power, the first phase will miss 20% of the true positives, and the second phase will miss 2%, for a combined power of 78%.

What about FPR, the False Positive Risk? Analyzing the Goldilocks version, the FPR of the experiment when using alpha1=0.01 is 7.1%. If the experiment is extended, an alpha2 of 0.046 results in an FPR of 17.5% for 98% power. The overall FPR is a linear combination of the two; if half of the statistically significant results can be declared after the first two weeks and half require two more weeks, then we have an FPR of 0.5*7.1%+0.5*17.5% = 12.3%, materially lower than the 22% mentioned at the end of Section 5.1. The calculations for the example here are in https://bit.ly/FalsePositivesInABTestsCalc, "FPR of two exp" tab. What is the impact to the agility of the organization? Not much. If 80% of experiments fail to be statistically significant improvements, then the extension will only be run for the portion of the 20% whose p-value is between 0.01 and 0.05. At Expedia, 42% of experiments will be extended (5.2%/(5.2%+7.2%)) based on Table 4, but even if conservatively assuming half, then 10% of experiments will have to be run for two more weeks. The average experiment duration will therefore be 15.4 days instead of 14 days.

6.1 Lowering the FPR

Bartoš and Maier [28] suggest that to reduce FPR, it is more efficient to reduce alpha than to increase power, a conclusion that aligns with our recommendation in Section 4.5.

If we set alpha2 to be 0.01, the second phase will have 92% power, thus we are slightly lowering the overall power. However, the FPR is lowered dramatically to 4.1%, so that the overall FPR is a linear combination of 7.1% and 4.1%. Assuming 50/50, the overall FPR is 0.5*7.1%+0.5*4.1% = 5.9%, materially lower than the above 12.3% and much lower than the original 22%. The calculations for the example here are in https://bit.ly/FalsePositivesInABTestsCalc "FPR of two exp" tab.

6.2 Increasing Power with Low FPR

Extending an experiment, as described above, reduces the FPR at a small cost to agility. What about extending an experiment to increase power? Suppose that instead of extending the experiment when the p-value is between 0.01 to 0.05, we extend it when the p-value is between 0.01 and 0.10?

The procedure is as follows:

1. Run the experiment with 80% power (e.g., two weeks).
2. If the p-value > 0.10, then stop. We can't reject the null.
3. If the p-value < 0.10, but the treatment effect estimate is negative, stop. The null is either true, or we're hurting users if it is false.
4. If the p-value < 0.01 and the treatment effect is positive, then stop. This is a win, as we can reject the null with low FPR.
5. Run phase 2 (replication run or extension), and reject the null as a win, only if the combined p-value < 0.01.

If our power for the first phase was previously 80%, we have now increased it to 88%. The first phase will miss only 12% of the true positives, and the second phase will miss an additional 8% for a combined power of 80%.

The change we made increases the number of experiments that go to phase 2, but not the rejection of the null for phase 1, so the FPR for the two phases do not change, having an overall FPR of 5.9%.

The impact here is to agility. Instead of 10% of experiments falling into the p-value range of 0.01 to 0.05, we might now have 20% in the range 0.01 to 0.10, and the average experiment duration will be 20% longer (e.g., 16.8 days instead of 14 days).

Note that we are extending experiments that are borderline positive, so the probability that we are exposing users to a statistically significant *negative* experience for longer is very low; it is much more likely that the treatment is "flat" with an effect close to zero than that it is statistically significantly negative. Gelman and Carlin [36] showed that the probability of a sign error is vanishingly small at 50% power, and much more so at 80% power recommended in this procedure.

To summarize, the new procedure keeps the same 80% power, but decreases the FPR from 22% to 5.9% for an increased average experiment time of 20%.

7 Success Rate of Ideas vs. Experiments

There is a lot of confusion in industry on the failure rate of ideas vs. experiments. We have heard people who quote success rate of experiments [14] and claim that 90% of ideas fail. While we believe that most ideas fail, the percentage is not as high as the failure

rate for experiments. The key to note is that an idea may be evaluated over multiple experiments. For example, think of the following realistic scenario:

8. An initial idea that starts off as an MVP (Minimum Viable Product/Feature) and an experiment is coded up and runs on one platform, say iOS. The first experiment launches and quickly aborted because of severe degradations to guardrail metrics. The bug is quickly found and fixed.
9. The second iteration starts, runs for two weeks, and is flat (not statistically significant). An analysis reveals that many users are aborting because they misunderstand the feature value. The designers change the UI.
10. A 3rd iteration starts, runs for two weeks, and is statistically significantly positive. Given the success, a decision is made to implement the idea for the other two platforms: Android and the Web.
11. The development for Android is delayed. A decision is made to test the Web version separately. The experiment needs to run for three weeks because it is a smaller segment than iOS. The results are flat. There are some hypotheses about whether the feature is less useful for web users, or whether the implementation could be improved, but a decision is made to wait for the results from the Android version.
12. Two weeks later, the Android version finishes development, and the experiment starts, but because the development was rushed, the experiment breaks a guardrail metric and is aborted. The bug is fixed.
13. The Android experiment is restarted and is run for three weeks, resulting in a statistically significant improvement.

A decision is made to launch the feature on all platforms. It is statistically significantly positive for iOS and Android, and flat for web, and consistency matters.

Summarizing, six experiments were run. Two were statistically significantly negative (a and e), two were flat (b and d), and two were statistically significantly positive (c and f). The "success rate" of experiments was 33%. The success rate of ideas was 100%, as we had one idea that ultimately launched.

Another factor worth noting is that the experiment success rate is likely to decline over time, as agility improves, and QA is reduced when the trust in the experimentation platform develops:

1. At Microsoft's early days, the success rate was 33% [41]. The trust in the early versions of the ExP platform was low (justifiably), and groups had high QA standards, so it was rare to see an experiment abort in the first day.
2. At Bing, the success rate was about 15% [18]. As the trust grew over time, the organization was taking more risks with less QA, because the experimentation platform provided a safety net. About 10% of experiments were aborted in the first day (usually in the first few hours).

There is obviously a tradeoff between increased agility and exposing users to software with increased bugs, but Microsoft Office had a ratio of 1:1 between developers and testers, which in an online world seems extreme when Amazon could do with a 10:1 ratio [53]. The value of increased agility and launching Minimum Viable Products/Features [54] was believed to be superior. After the successful experience with increased agility and reduced QA at Bing, Satya Nadella reduced QA across Microsoft when he became CEO [55].

8 SUMMARY

There has been an ongoing debate in the software industry, with some claims that we should increase the alpha threshold for accepting stat-sig results from 0.05, or run one-tail tests, because 0.05 is too stringent. In the extreme, Longden [56] wrote that "In certain situations, there is no reason why 70% significance (alpha=0.3) isn't a completely acceptable reflection of a risk approach." This is a common misinterpretation of p-values [57; 25; 26; 27]. The False Positive Risk is a much more intuitive metric, highlighting the ratio of incorrect statistically significant results.

The key question is the success rate for an organization, and we proposed several approaches to estimate it, with examples from Optimizely and Expedia. Whether you weigh false positives to negatives as 1-to-1 or 3-to-1, most organizations should lower their alpha to reduce the loss stemming from false negatives and false positives. Expedia, which has used an alpha of 0.1, is now in the process of lowering the alpha given these results.

In our analysis, we assumed a fixed success rate across all experiments. In practice, this is a classical bias-variance tradeoff question; if there is reason to believe that different groups or organizations have different success rates or risk profiles, then with enough data it makes sense to fragment the data and analyze them separately.

We proposed a modified procedure for experimentation, based in sequential group testing, that selectively extends experiments to reduce false positives, increase power, at a small increase to runtime.

We concluded with a discussion of the difference between ideas and experiments in practice, terms that are often incorrectly used interchangeably. While it is likely that most ideas fail, it is at a lower rate than the median 90% rate for experiments.

ACKNOWLEDGMENTS

We thank Alex Deng and Lukas Vermeer for feedback on a draft of this paper, and Andrew Gelman, Adam Gustafson, and Art Owen for early discussions on ideas in this paper. The authors acknowledge Cristina McGuire and Connor Culleton for analysis

and discussion around false discovery risk analysis at Expedia Group.

REFERENCES

[1] **Kohavi, Ron, Tang, Diane and Xu, Ya.** *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* s.l. : Cambridge University Press, 2020.

[2] **Kohavi, Ron and Longbotham, Roger.** Online Controlled Experiments and A/B Tests. [book auth.] Dinh Phung, Geoffrey I Webb and Claude Sammut. *Encyclopedia of Machine Learning and Data Science.* New York, NY : springer, 2023.

[3] **Luca, Michael and Bazerman, Max H.** *The Power of Experiments: Decision Making in a Data-Driven World.* s.l. : The MIT Press, 2020.

[4] **Thomke, Stefan H.** *Experimentation Works: The Surprising Power of Business Experiments.* s.l. : Harvard Business Review Press, 2020.

[5] **Georgiev, Georgi Zdravkov.** *Statistical Methods in Online A/B Testing: Statistics for data-driven business decisions and risk management in e-commerce.* s.l. : Independently published, 2019. 978-1694079725.

[6] **Montgomery, Douglas C.** *Design and Analysis of Experiments.* 10th edition. s.l. : Wiley, 2019.

[7] **Imbens, Guido W and Rubin, Donald B.** *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* s.l. : Cambridge University Press, 2015. 978-0521885881.

[8] **Box, George E.P., Hunter, J Stuart and Hunter, William G.** *Statistics for Experimenters: Design, Innovation, and Discovery.* 2nd. s.l. : John Wiley & Sons, Inc, 2005. 0471718130.

[9] **Gerber, Alan S and Green, Donald P.** *Field Experiments: Design, Analysis, and Interpretation.* s.l. : W. W. Norton & Company, 2012. 978-0393979954.

[10] *The Surprising Power of Online Experiments: Getting the most out of A/B and other controlled tests.* **Kohavi, Ron and Thomke, Stefan.** Sept-October, 2017, Harvard Business Review, pp. 74-92.

[11] **Kohavi, Ron, et al.** Online Controlled Experiments at Large Scale. *KDD 2013: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2013. http://bit.ly/ExPScale.

[12] *Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology.* **Larsen, Nicholas, et al.** 2, s.l. : The American Statistician, 2023, Vol. 78.

[13] *Top Challenges from the first Practical Online Controlled Experiments Summit.* **Gupta, Somit, et al.** 1, June 2019, Vol. 21.

[14] *A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments.* **Kohavi, Ron, Deng, Alex and Vermeer, Lukas.** Washington DC, USA : ACM, New York, NY, USA, 2022. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22).

[15] *The reproducibility of research and the misinterpretation of p-values.* **Colquhoun, David.** 4, 2017, Royal Society Open Science.

[16] **Wacholder, Sholom, et al.** Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. *Journal of the National Cancer Institute.* 2004, Vol. 96, 6. http://jnci.oxfordjournals.org/content/96/6/434.long.

[17] *Why Most Published Research Findings Are False.* **Ioannidis, John P.** 8, 2005, PLoS Medicine, Vol. 2, p. e124. http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124.

[18] **Kohavi, Ron, et al.** Seven Rules of Thumb for Web Site. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14).* 2014.

[19] *Causality and Statistical Learning.* **Gelman, Andrew.** 3, s.l. : American Journal of Sociology, 2011, Vol. 117.

[20] *Estimating the Reproducibility of Psychological Science.* **Open Science Collaboration.** 6251, 2015, Science, Vol. 349.

[21] *Ego depletion: is the active self a limited resource?* **Baumeister, Roy F, et al.** 5, May 1998, Journal of personality and social psychology, Vol. 74, pp. 1252-265.

[22] **Schimmack, Ulrich.** Replicability Report No. 1: Is Ego-Depletion a Replicable Effect? *replicability-Index.* [Online] April 18, 2016. https://replicationindex.com/2016/04/18/is-replicability-report-ego-depletionreplicability-report-of-165-ego-depletion-articles/.

[23] *A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect.* **Vohs, Kathleen D, et al.** 10, 2021, Psychological Science, Vol. 32, pp. 1566-1581.

[24] *The "File Drawer Problem" and Tolerance for Null Results.* **Rosenthal, Robert.** 3, 1979, Psychological Bulletin, Vol. 86, pp. 638-641.

[25] *A Dirty Dozen: Twelve P-Value Misconceptions.* **Goodman, Steven.** 2008. Seminars in Hematology.

[26] *Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.* **Greenland, Sander, et al.** 2016, European Journal of Epidemiology, Vol. 31, pp. 337-350.

[27] **Vickers, Andrew J.** *What is a p-value anyway? 34 Stories to Help You Actually Understand Statistics.* s.l. : Pearson, 2009. 978-0321629302.

[28] *Power or Alpha? The Better Way of Decreasing the False Discovery Rate.* **Bartoš, František and Maier, Maximilian.** Nov 08, 2022, Meta-Psychology, Vol. 6.

[29] *Z-curve 2.0: Estimating Replication Rates and Discovery Rates.* **Bartoš, František and Schimmack, Ulrich .** 2022, Meta-Psychology, Vol. 6.

[30] *Redefine Statistical Significance.* **Benjamin, Daniel J., et al.** 1, September 1, 2017, Nature Human Behaviour, Vol. 2, pp. 6-10.

[31] **Optimizely.** Confidence intervals and improvement intervals. *Optimizely.* [Online] 2023. https://support.optimizely.com/hc/en-us/articles/4410283895821-Confidence-intervals-and-improvement-intervals.

[32] **Georgiev, Georgi.** One-tailed vs Two-tailed Tests of Significance in A/B Testing. *Analytics Toolkit.* [Online] August 8, 2018. https://blog.analytics-toolkit.com/2017/one-tailed-two-tailed-tests-significance-ab-testing/.

[33] **Skotara, Nils.** Raising the bar by lowering the bound. *Booking.ai.* [Online] Nov 1, 2023. https://booking.ai/raising-the-bar-by-lowering-the-bound-3b12d3bd43a3.

[34] **Gabster, Elizabeth, et al.** *Evolution of Experimentation.* 2023.

[35] **Reinhart, Alex.** *Statistics Done Wrong: The Woefully Complete Guide.* s.l. : No Starch Press, 2015. 978-1593276201.

[36] *Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.* **Gelman, Andrew and Carlin, John.** 6, 2014, Perspectives on Psychological Science, Vol. 9, pp. 641 –651.

[37] **Kohavi, Ron.** *Why positive A/B test results should always be given a haircut.* Dec 3, 2023.

[38] *Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments.* **Lee, Minyong R and Shen, Milan.** London : ACM, 2018. KDD 2018: The 24th ACM Conference on Knowledge Discovery and Data Mining.

[39] *Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data.* **Zöllner, Sebastian and Pritchard, Jonathan K.** 4, 2007, The American Journal of Human Genetics, Vol. 80, pp. 605-615.

[40] *On Post-Selection Inference in A/B Tests.* **Deng, Alex, et al.** 2021. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 2743-2752.

[41] **Kohavi, Ron, Crook, Thomas and Longbotham, Roger.** Online Experimentation at Microsoft. [ed.] Peter van der putten, Gabor Melli and Brendan Kitts. *Third Workshop on Data Mining Case Studies and Practice Prize.* 2009. http://www.appliedaisystems.com/papers/DMCS2009_Workshop proceedings4.pdf.

[42] **Wikipedia contributors.** Fisher's method. *Wikipedia.* [Online] Dec 2023. http://en.wikipedia.org/wiki/Fisher%27s_method.

[43] **Kohavi, Ron.** *Meta Analysis Spreadsheet.* 2021.

[44] *Publication Decisions Revisted: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa.* **Sterling, T D, Rosenbaum, W L and Weinkam, J J.** 1, 1995, The American Statistician, Vol. 49, pp. 108-112. https://gwern.net/doc/statistics/bias/1995-sterling.pdf.

[45] *A/B Testing with Fat Tails.* **Azevedo, Eduardo M., et al.** 12, 2020, Journal of Political Economy, Vol. 128.

[46] **Georgiev, Georgi.** *What Can Be Learned From 1,001 A/B Tests?* Oct 17, 2022.

[47] **Casella, George and Berger, Roger.** *Statistical Inference.* 2nd. 2002.

[48] *Statistical Inference in Two-Stage Online Controlled Experiments with Treatment Selection and Validation.* **Deng, Alex, Li, Tianxi and Guo, Yu.** Seoul, Korea : International World Wide Web Conference (IW3C2), 2014.

[49] **DeMets, David L. and Lan, Gordon.** The alpha spending function approach to interim data analyses. [book auth.] P.F. Thall. *Recent Advances in Clinical Trial Design and Analysis.* s.l. : Springer, 1995.

[50] *Discrete sequential boundaries for clinical trials.* **Lan, K.K. Gordon and DeMets, David L.** 3, 1983, Biometrika, Vol. 70, pp. 659-663.

[51] *Group sequential methods in the design and analysis of clinical trials.* **Pocock, Stuart J.** 2, Aug 1977, Biometrika, Vol. 64, pp. 191-199.

[52] *A Multiple Testing Procedure for Clinical Trials.* **O'Brien, Peter C. and Fleming, Thomas R.** 3, September 1979, Biometrics, Vol. 35, pp. 549-556.

[53] **Arguelles, Carlos.** The Paradigm Shifts with Different Dev:Test Ratios. *Medium.* Aug 31, 2021.

[54] **Ries, Eric.** *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* . s.l. : Crown Business, 2011. 978-0307887894.

[55] **Patrizio, Andy.** Would Microsoft really cut its QA department? *Network World.* July 15, 2014.

[56] **Longden, Jonny.** *The Power of Experimentation: A/B testing for startups and low traffic websites.* February 2023.

[57] **Clayton, Aubrey.** *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science.* s.l. : Columbia University Press, 2021. 0231199945.

Appendix A: FPR for p-value equals case

Colquhoun [15] provides an online calculator at http://fpr-calc.ucl.ac.uk/. Plugging in an example of an e-commerce site with 5% conversion rate and MDE of relative 5%, and using the Optimizely success rate and alpha of 0.10 [34], we have the following:

1. P-value of 0.05 (single tail equivalent to 0.10 two-tailed).
2. Probability of real effect: 9.3%, based on 12% observed win rate for Optimizely's report.
3. Samples required for 80% power:

   16*(0.05*0.95)/(0.05*0.05)^2 = 121,600
4. Effect size as multiple of MDE:

   0.05*0.05/sqrt(0.05*0.95)= 0.0115

(Note that the items 3 and 4 above are only necessary for the calculator to end up with 80% power. You would get the same result if you input n=16 and effect size of 1, which Colquhoun uses in his paper.)

The simulation shows an FPR for p-less-than-case of 37.6%, very close to our Bayes Rules estimate of 37.8%. The p-equals case is 80.7% (image below).

An experiment with a p-value of 0.10 (90% confidence) for a typical Optimizely has over 80% probability of being a false positive!