Optimizely

A Practical Guide to

Statistics for Online Experiments

How anyone can think and act like a statistician

Hello,

This is Pete Koomen. Just a quick thank-you for downloading this guide and taking the time to dive into a topic that we're passionate about here at Optimizely.

We've spent the past few years building a company and a product that enables our customers to turn data into action. We want to make it possible for anyone, in any company, to use A/B testing and optimization as processes that will help them make decisions, reach their conversion goals, and transform their business.

Although we know you value data and hard facts when growing your business, you make intuition-driven decisions about your results. To make sure you make the best decisions, we're committed to giving you the very best data. This means that the statistical underpinnings of our platform need to evolve to keep pace with how you want to use your results.

In this guide, we're going to take things one step further. We'll cover the essential 'Stats IQ' topics you need to take your A/B testing and optimization efforts to the next level. We'll show you how to trust your test results, make recommendations, and get to significance more confidently than ever before.

We hope you'll take some of these concepts back to your company, clients, and teammates to share them and help them to make informed decisions and get great test results.

Cheers,

Pete Koomen
Chief Technical Officer, Optimizely

Table of Contents

1. Why we need statistics
2. How statistics have been traditionally used
3. How statistics have (or haven't) adapted for the online world
   > Misunderstanding of statistics leads to errors
   > Calculating sample size
4. What you need to know about statistical error
5. How Optimizely is creating an always-valid results calculation
   > Continuous monitoring error
   > Multiple comparisons error
6. Tips for running a statistically sound experiment
7. How to communicate experiment results
   > Confidence intervals
8. How to set your statistical significance threshold
9. How to reach statistical significance if you have low traffic
10. Statistical terms glossary

Why we need statistics

Statistics are the underpinning of how Optimizely's customers use data to make decisions. We use experiments to understand how changes we make affect the performance of our online experiences. To do this, we need a framework for evaluating how likely those changes are to have an impact, positive or negative, on a business over time.

To run great experiments, investing in an understanding of statistics is one of the most important skills you can develop. Statistics provide inference on your results and help to determine whether you have a winning variation. Using statistical values to decide leads to stable, replicable results you can bet your business on. Lack of understanding can lead to errors and unreliable outcomes from your experiments.

We're committed to making statistics evolve to fit the way you run experiments—in online, real-time environment with new data coming in every second. This guide will outline some of the core concepts behind your results and how to run statistically sound experiments. Some of these tips will be specific to Optimizely, but most will be relevant to any testing platform.

STATISTICAL TERMS TO KNOW

Statistical significance
The number in an experiment results report that represents the likelihood that the difference in conversion rates between a given variation and the original (the effect) is not due to chance. This is displayed as a value between 0 and 99%, or 1 minus the false discovery rate. (We'll discuss these concepts in greater detail later.)

Traditional statistics
This is how we'll refer to statistics that were developed to support experimentation in offline situations. These methods are still used by many online A/B testing platforms to calculate statistical significance. They are also called traditional hypothesis testing or fixed-horizon testing.

Sample size
The number of participants in an experiment that are required to achieve a statistically significant result. Typically, this value is calculated in traditional statistical experiments with the help of a sample size calculator.

Effect
The difference in conversion rate between a variation treatment and control in an experiment.

Statistical error
A result that reaches statistical significance that does not represent a significant result. This is effectively the inverse of statistical significance. If an experiment variation reaches 95% statistical significance, for example, there is a 1 in 20 chance that the effect was found because of spurious trends in the experiment data.

For a full list of terms referenced in this guide (and some that weren't), see Section 10.

How statistics have traditionally been used

The use case for statistics in digital experiments is extremely different from the world in which traditional statistics were conceived. Statistical methods were first applied to experiments in the fifteenth century, and have been used scenarios that include:

AGRICULTURE
Farmers planted a control and variation strain of crops to determine which seed variety produced a better crop.

MEDICINE
Researchers administer an experimental drug treatment with mice against a placebo control group to determine if the medicine has any effect.

OTHERS,
like economics, engineering, behavioral research, and more.

In these experiment scenarios, analyzing experiment data occurs at one point in time. The farmer evaluates the difference in crop yield at harvest, a researcher determines whether their drug was effective once the full course of the medicine is administered.

The process of collecting data for analysis at one point in time is also known as a **fixed horizon**. At this point, a p-value is typically calculated and the experimenter determines whether there is a statistically significant difference between the variation and control groups.

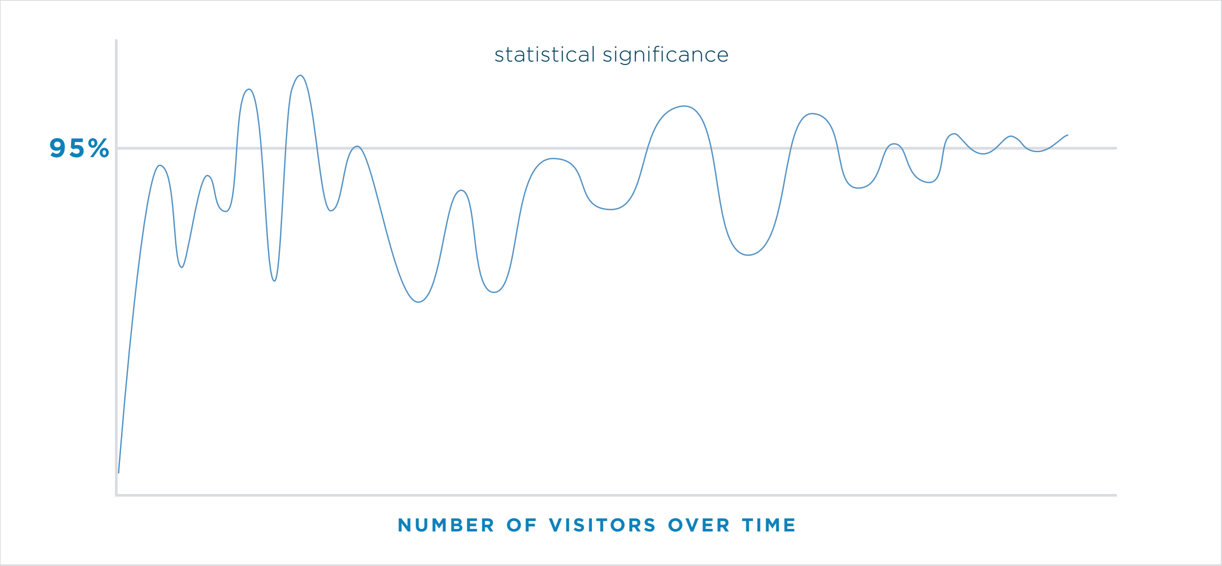
Of course, experiments are no longer confined to offline environments. This means that the considerations for how statistical significance is calculated and used should change. Let’s discuss how statistics have and haven’t changed, and how experimenters A/B testing online are using them.

How statistics have (or haven't) adapted for the online world

Once technology adapted to support running experiments in online environments, statistics were copied over from traditional methods to calculate the results. This is how many A/B testing solutions currently calculate statistical significance.

Misunderstanding of statistics leads to errors

Unfortunately, A/B testing platforms don't wait until a fixed horizon to display statistical significance. Results are calculated and updated live, as data is collected. This means that statistical significance can change and vary widely as more visitors participate in an experiment. In many cases, statistical significance can waver below and above a significance threshold over time, like this:



This is frustrating for experimenters; since they may watch insignificant results become significant, but also waver in the opposite direction. The changing declaration makes it difficult to call an experiment result with confidence.

More often than not, when you run an experiment, it is tempting to check results continuously and look for a discovery. It is also very tempting to stop an experiment when it reaches statistical significance the first time, even if it is before the needed sample size for that effect.

This is an unsafe method of conducting an experiment.

Sometimes called the "peeking analyst" problem, or continuous monitoring, this problem represents a problem with traditional statistics. They're just not built for online experiments; they are difficult to apply correctly, and they require extra precautions (like calculating sample size) that just aren't realistic for how businesses want to run their experiments.

The good news is that there's actually a pretty simple yet elegant statistical solution that lets you see results that are always valid. It's called sequential testing, and we'll discuss it in Section 5 below.

Calculating sample size: the calculator approach

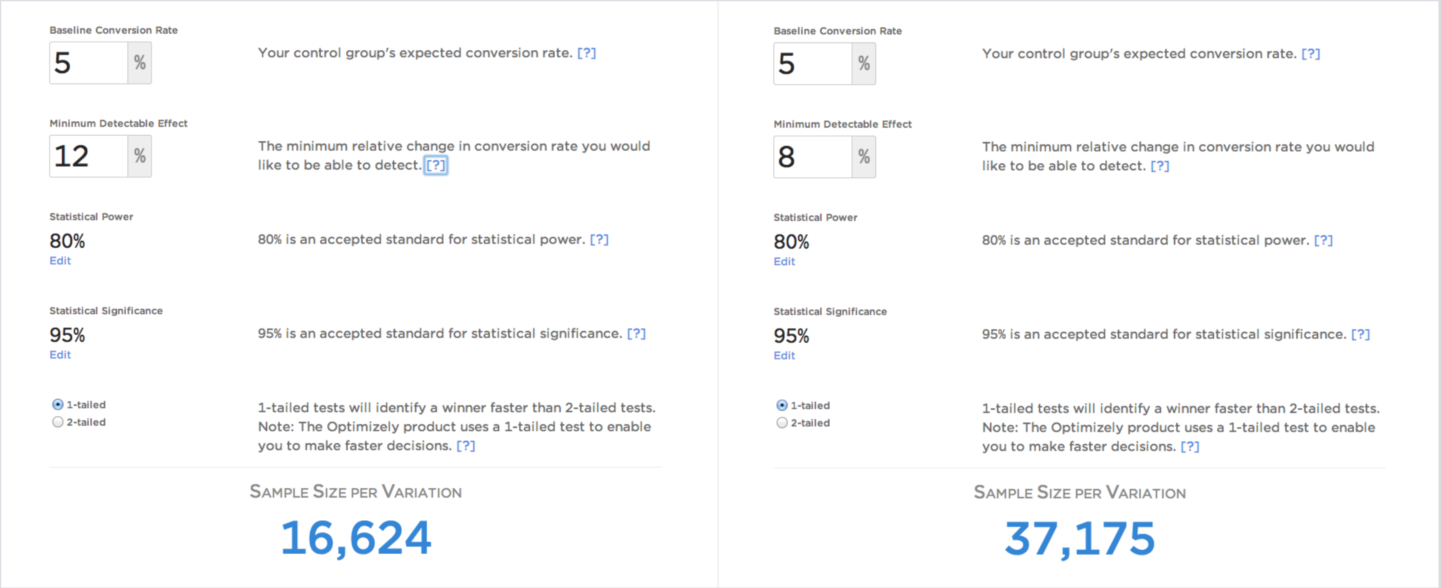
If you run experiments with an in-house solution or tool that uses traditional statistics to calculate results, you should use a sample size calculator to prevent finding a false result.

This ensures that the experiment has adequate data to ensure a statistically significant result, if there is one to be found.

To set a sample size, you need to have your baseline conversion rate handy. You also need to make your best guess about the Minimum Detectable Effect, or expected conversion rate lift, you want to see from your test.

These two numbers will help to predict the needed sample size for that improvement at a given statistical significance and statistical power.*

* Sometimes expressed as (1 - type II error), this is the probability that an experimenter will detect a difference when it exists. It is also the probability of correctly rejecting a null hypothesis.



Baseline Conversion Rate
5 %
Your control group's expected conversion rate. [?]

Minimum Detectable Effect
12 %
The minimum relative change in conversion rate you would like to be able to detect. [?]

Statistical Power
80%
Edit
80% is an accepted standard for statistical power. [?]

Statistical Significance
95%
Edit
95% is an accepted standard for statistical significance. [?]

● 1-tailed
○ 2-tailed
1-tailed tests will identify a winner faster than 2-tailed tests. Note: The Optimizely product uses a 1-tailed test to enable you to make faster decisions. [?]

SAMPLE SIZE PER VARIATION
16,624

Baseline Conversion Rate
5 %
Your control group's expected conversion rate. [?]

Minimum Detectable Effect
8 %
The minimum relative change in conversion rate you would like to be able to detect. [?]

Statistical Power
80%
Edit
80% is an accepted standard for statistical power. [?]

Statistical Significance
95%
Edit
95% is an accepted standard for statistical significance. [?]

● 1-tailed
○ 2-tailed
1-tailed tests will identify a winner faster than 2-tailed tests. Note: The Optimizely product uses a 1-tailed test to enable you to make faster decisions. [?]

SAMPLE SIZE PER VARIATION
37,175

Changing the predicted minimum detectable effect dramatically changes the required sample size for an experiment

The drawback to this approach is that it is inefficient. With a smalle effect, you have to wait for a large sample size to find significant results. Set a larger effect, and you risk missing out on smaller improvements.

In reality, you just don't know what effect a variation might have, so committing in advance to a hypothetical lift just doesn't make a lot of sense, especially when you have data coming in to your experiment constantly.

Choosing a sample size and sticking to it, though imperfect, is the best way to ensure that your results are statistically valid and that you've avoided statistical errors if you're using a platform powered by traditional statistics.

What you need to know about statistical error

A statistical error is a result that reaches statistical significance that does not represent a significant result.* Statistical errors happen because of spurious runs of experiment data that paint a misleading picture of what's actually happening with your visitors and users. Sometimes referred to as **false positives** or **Type I errors**, these are misleading signals from your experiment that won't translate into true improvements over time.

What you need to know about error is how to quantify it and how to control for it. Because your experiment data will always rely on data that has some degree of uncertainty (two runs of an A/B testing will practically never be the same), statistical errors are bound to happen. Understanding what they are and how to think about them is a key step to getting better results.

Error rate is the chance that your experiment incorrectly showed a change between the original and variation with a conclusive level of statistical significance. This is effectively the inverse of statistical significance. This is how you can quantify your risk of finding a result in error:

If an experiment variation reaches 95% statistical significance, there is a 1 in 20 chance that the effect was found due to random chance, and not because of a change you introduced.

Thinking about statistical significance in terms of error can often change the way experimenters want to run tests and at what point they'll take action on the results. For more examples of scenarios where different levels of statistical significance may be appropriate, check out Section 8**.**

WHAT YOU NEED TO KNOW ABOUT STATISTICAL ERROR

- It's an inevitable part of every experiment.
- Carefully choose the statistical significance threshold you set for your experiments.
- Share your knowledge about errors with your teammates involved in testing, so they are aware of the implications on your experiment results.

Statistics are evolving to make random errors in your experiment data less likely. See Section 5 for more details.

* For the sake of clarity, we are primarily addressing Type I statistical errors. These are errors that surface significant experiment results when there actually are none. Type II statistical errors are errors that prevent significant results from surfacing, and can also be read as 1 minus statistical power. Type I errors because they present a higher risk that an experimenter may see and take action on an incorrect result.

If you've ever implemented a winning variation from an experiment and not seen that lift appear, it is possible that you took action on a statistical error. At best, maybe the change had no effect, but at worst, it could have negatively impacted your conversion rate. It's essential to choose a method of calculating results that will help keep errors where you expect them to be, and to set the appropriate threshold for significance for your organization.

How Optimizely is creating an always-valid statistical significance calculation

Statistical name: Continuous monitoring

In traditional statistics, statistical significance is calculated assuming that the test will run over a set sample size and will conclude immediately after that sample size is reached. Statisticians call this a **fixed horizon.**

Optimization platforms calculate statistical significance in real-time, but because they use a fixed horizon methodology, they assume you will only make a decision on results at the final sample size, and won't collect any more data, or check results early.

If your testing platform is set to accept a 5% chance of error (a 95% statistical significance level), it's essentially using all of its 'error budget' on one predetermined sample size. Making decisions at any other time puts the experiment over budget!

Because of this behavior, there is a much higher likelihood of seeing a false statistically significant result. Statisticians call this **type I error**—detecting an effect that is not actually present in your results.

Since you should have be able to use experiment results as they are collected without being hampered by sample size, Optimizely has introduced sequenstial testing,

With **sequential testing,** your statistical significance now increases over time as your experiment collects more data, just like it should. This lets you check your results in real time and know you're seeing a valid statistical determination whenever you look.

Because you don't need to set a sample size or detectable effect in advance, sequential testing also allows you to test more accurately without sacrificing speed. You're able to act on the effect found in the experiment, instead of committing to it prior to starting, which is unintuitive and can be too conservative.

Statistical name: Multiple comparisons error

Another pitfall of using traditional statistics for optimizing your website involves testing lots of goals and variations at once. When you do this, you're increasing the chance that you'll find false positive results hidden among your significant results.

As an experimenter, you will make find many winning and losing goals. Your job is to maximize the number of these winners that are correct, or are true discoveries. Usually, this is achieved by setting a significance level of 95% and therefore only accepting 5% false positive rate in test results. However, this is a deceptive cutoff in tests with multiple comparisons of goals and variations.

Maximizing the number of correct calls on goals and variations is actually not the same as minimizing the probability that you'll find significant results where none actually exist (**false positives**.) Each additional variation and goal adds a new combination of individual comparisons to an experiment. In a scenario where there are four variations and four goals, that's 16 potential outcomes that need to be controlled for separately.

Most A/B testing platforms calculate statistical significance for individual goal-variation combinations, without taking into account the other comparisons being made. This creates a big impact on the likelihood of false discoveries, which isn't reflected by statistical significance.

Multiple testing correction applies corrections for the number of goals and variations in an experiment, drastically reducing the chance of false positives among your significant results. With this technique, you can test as many goals and variations as you want without worrying about compromising the statistical validity of your results.

Tips for running a statistically sound experiment

In any experiment you, need to ensure that the data gathered from the experiment reflects the effect of only the variable(s) being tested, avoiding errors that are introduced from flaws in experiment administration.

Here's a pre-experiment checklist to make sure that you're running a fair and accurate test every single time:

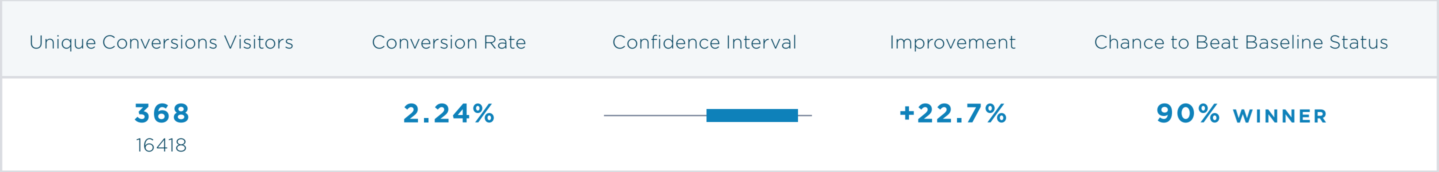
POTENTIAL ERROR SCENARIO	RECOMMENDATIONS
□ Correctly implement your A/B testing software. Introducing a bug into your experiment or miscounting experiment participants can skew results.	• Use a diagnostics report to validate that your software is correctly installed. • Run an A/A test (With caution; read more about A/A tests here.)
□ QA your test before it starts. Ensure that any difference between control and variation is by design only.	• Treat an experiment variation as if it were a new feature or lines of code for your website. • Force a variation to display on a selection of browsers and devices before you set the experiment live.
□ Check for seasonal effects on visitor behavior. Behavior might be inconsistent different days of the week, during a promotional period, or during a holiday season.	• Validate experiment results during periods where seasonal effects in your industry are less likely. • Try to segment your results and understand variations in behavior between segments (new versus returning traffic, for instance). • Try to avoid running weekday-only or weekend-only experiments, unless you are doing so intentionally.
□ Check for multiple tests running on the same page. If your visitors participate in multiple experiments, you might affect behavior in untrackable ways.	• Run a single test per page, and be aware of downstream effects of an A/B test if you are running multiple tests within the same funnel or click path. • Use multivariate tests when modifying multiple variables on one page to maintain a focus on causal variables.
□ Run your variation(s) unchanged from start to finish. Making changes to a variation while an experiment is paused or running can show different variations to the same group of visitors.	• Don't change a variation mid-experiment. • If you need to make the change, scrap the first experiment and start a new one.
□ Plan your traffic allocation before starting (50-50, 90-10, or some other distribution.) Changing a traffic allocation mid-experiment changes the number of visitors that see each variation, but your testing software will display average metrics.	• It's best not to change your traffic allocation. • If you change your setting mid-experiment, re-calculate your metrics for periods where the allocation remains constant, and consider the experiment results separately.

How to communicate experiment results

Setting proper expectations around experiment results and the outcomes of implementing a winning variation are essential for the long-term success of an optimization program.

Provide the appropriate level of context for your audience. We recommend that all results that are shared should include the following:

- **Overview of the experiment:** What the test was, on which pages it was run, to what types of visitors, and what the goals measured were.
- **Screenshots:** Capture the original and variation(s.)
- **Experiment hypothesis:** Was this a new hypothesis, or a modification of a hypothesis from a previous experiment?
- **The results:** The total number of visitors that participated in the experiment, timeframe, and the observed effect.
- **The statistical significance:** Include both the significance and the confidence interval.



Unique Conversions Visitors	Conversion Rate	Confidence Interval	Improvement	Chance to Beat Baseline Status
368 16418	2.24%		+22.7%	90% WINNER

Statistical significance and confidence interval, reported in tandem, provide a clear and accurate view of how likely it is that the experiment results will manifest once they are implemented.

As we've discussed, statistical significance is the likelihood that the experiment results were found because of a variable you introduced, and not because of random chance.

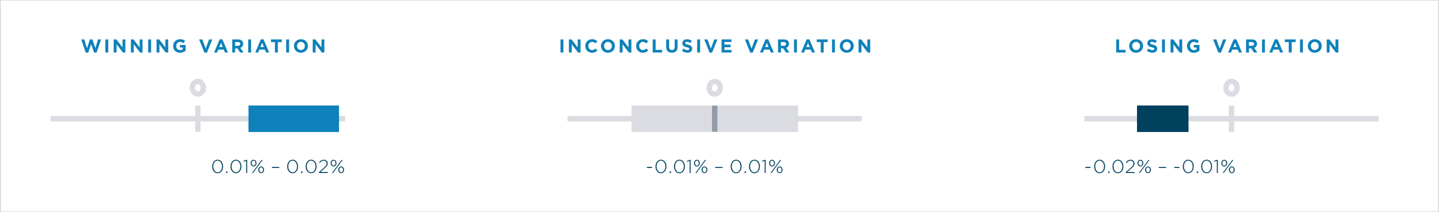
Confidence intervals

When applied to A/B testing, confidence intervals represent the likely range of the improvement in an experiment.

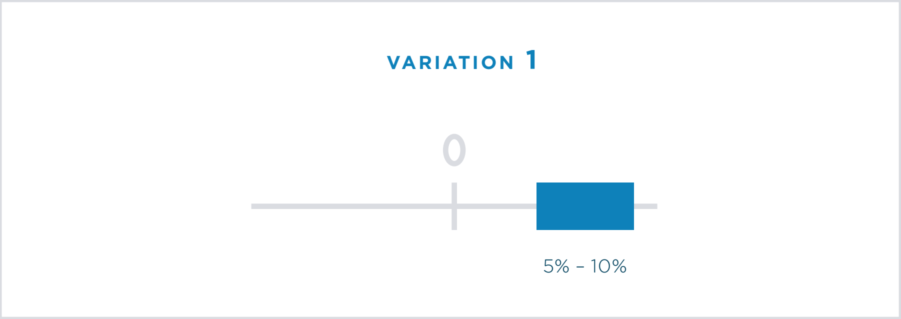
Statistical significance tells you whether a variation is outperforming or underperforming the baseline, at some level of confidence. Confidence intervals tell you the *magnitude* of the difference between the two conversion rates.

Confidence intervals represent a range of values for which the absolute difference in conversion rates are likely to be found. When a winning variation reaches statistical significance, its confidence interval lies entirely above 0%. A statistically significant losing variation would lie entirely below zero.

In Optimizely, your confidence interval is set at the same level that you set your statistical significance threshold for the project. So if you accept 90% significance to declare a winner, you also accept 90% confidence that the interval is accurate.



A winning variation will have a confidence interval that is completely above 0%. An inconclusive variation will have a confidence interval that includes 0%. A losing variation will have a confidence interval that is completely below 0%.



In the example above, Variation #1's confidence interval goes from 5% to 10%. This means there is a 90% chance that the true difference between variation #`1 and the baseline is between 5% and 10%. Note that these differences are absolute, not relative. In other words, a 5% absolute difference from 10% is 5-15%, not 9.5-10.5%.

How to set your statistical significance threshold

Setting the appropriate level of statistical significance for your experiments is closely tied to how willing you are to tolerate statistical errors in your experiments.

The widely accepted standard for statistical significance is 95%. However, it is possible to take action on your results at either a higher or lower level of statistical significance.

Different scenarios and organizations might require different levels of statistical significance in running experiments. Since A/B tests can be used for everything from validating a product concept to testing checkout flows to increasing revenue per visitor, statistical significance and error can be considered differently in each scenario

Since statistical significance threshold can vary by situation, Optimizely supports statistical significance as a project-level setting. Adjust your threshold up or down to meet your optimization needs.

Here are the factors you should consider when assessing the amount of risk you can tolerate in an experiment:

- **AVAILABLE TIME AND TRAFFIC TO RUN THE EXPERIMENT**
  Higher levels of statistical significance will require more data, and will thus need to run longer. If you want to run more experiments, you will either need to aim for larger effects in designing your experiments (see Section X.X for details), or settle for lower levels of statistical significance.

- **EXPECTED EXPERIMENT ROI**
  Higher projected ROI can lead to increased expectations in within your team or from an executive sponsor. Consider running these experiments longer to achieve higher levels of statistical significance.

  If the experiment is not focused on ROI, but instead concept validation, you may be willing to accept a higher level of statistical error for the sake of moving quickly and validating or disproving a hypothesis.

- **EXPECTED ENGINEERING INVESTMENT**
  To be sure that you have found a true winner before asking for engineering resources, allow your experiments to run to higher levels of statistical significance.

  The opposite outcome is still valuable; finding an inconclusive or losing variation avoids spending valuable resources on a change that may not have translated into an improvement.

- **PRIMARY EXPERIMENT GOAL**
  Is the conversion goal a purchase, engagement, reducing a page bounce rate, or another metric?

  You may choose to assign different required levels of significance to different types of goals based on how much error you could risk with each of them.

Expert tip: Repeated testing

In a situation where you cannot risk a statistical error, you can put your experiment to the test by running it repeatedly. This can help to ensure that you did not encounter a statistical error by random chance. Try for a best-of-three outcome to determine whether or not an effect was truly there. This strategy may only be possible for organizations that have time, traffic, and resources available to run the same test multiple times, but it is the absolute safest method for avoiding statistical errors.

However, be sure to account for complications that can arise from repeated testing: underlying changes in visitor behavior due to experiments being run at different points in time is one consideration. Be careful not to run multiple tests on the same page or in the same funnel to avoid interference effects.

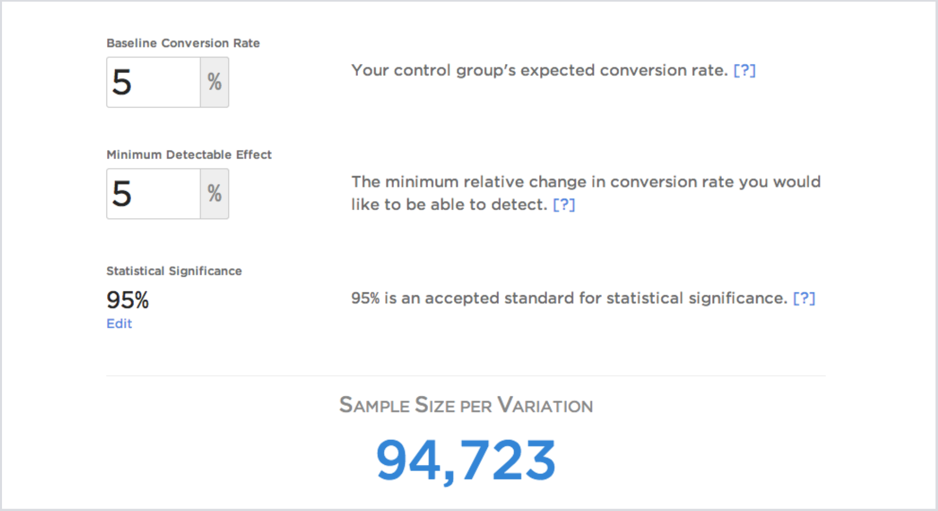
How to reach statistical significance if you have low traffic

A frequent challenge that many organizations face when planning their experiments is the issue of low traffic or low conversions. Experiments on low-traffic pages can run for extended periods of time to reach optimal levels of statistical significance. Since there is an opportunity cost to always having an experiment running, many organizations will forgo running experiments altogether.

Fortunately, you can run a successful optimization program without high traffic. Low-traffic experiences can be tested with adjusted expectations of what can and should be tested to reach statistical significance.

To reach statistically significant results on low-traffic pages, the variation must create a relatively large effect against the original. If the difference in conversion rate is substantially different, the statistical conclusion that the variation and control are different becomes evident much faster.

How do we quantify "substantially different"? In this scenario, a sample size calculator is helpful for visualizing the interdependence of effect and sample size.



Baseline Conversion Rate
5 %
Your control group's expected conversion rate. [?]

Minimum Detectable Effect
5 %
The minimum relative change in conversion rate you would like to be able to detect. [?]

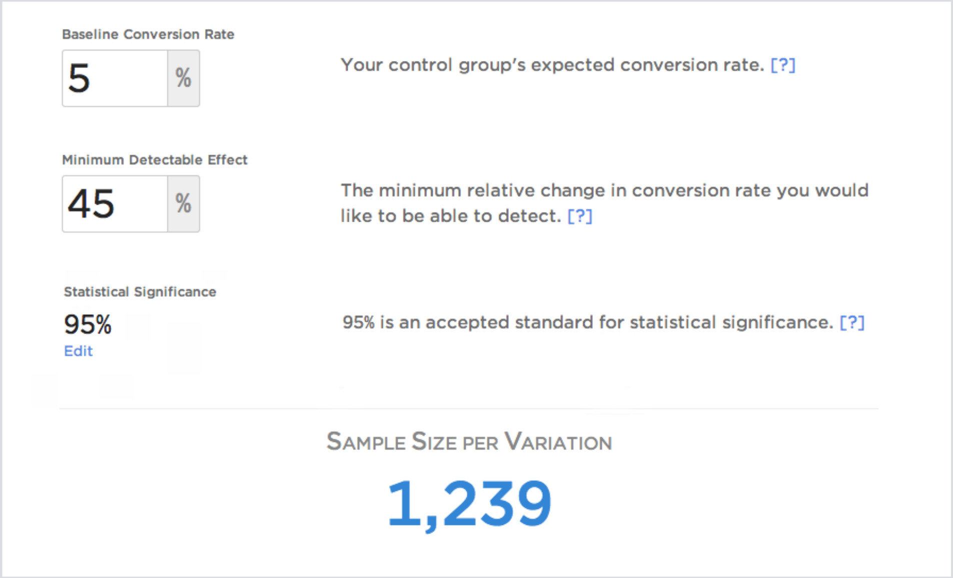
Statistical Significance
95%
Edit
95% is an accepted standard for statistical significance. [?]

SAMPLE SIZE PER VARIATION
94,723

Given a 5% baseline conversion rate, you need 94,723 unique visits per branch, totaling 189,448 to detect a 5% relative effect in the conversion rate.

Teams with low traffic would have to wait weeks, months, or longer to detect this small effect on a low-conversion. Subtle effects in the 5% range are off the table.

What about a larger effect?



Baseline Conversion Rate

5 %

Your control group's expected conversion rate. [?]

Minimum Detectable Effect

45 %

The minimum relative change in conversion rate you would like to be able to detect. [?]

Statistical Significance

95%

Edit

95% is an accepted standard for statistical significance. [?]

SAMPLE SIZE PER VARIATION

1,239

Given the same 5% baseline conversion rate, you need 1,239 visits for a 45% minimum detectable effect (MDE.) This change is detectable at a more reasonable pace than the small effect above.

The sample size reduces further if you're working with a higher conversion rate. With a 10% baseline conversion rate, it can detect a 45% difference with 485 visits—less than half the amount required for a 5% baseline conversion rate.

Effect, or MDE, and the baseline conversion rate have a negative relationship with sample size. When the baseline conversion rate increases, the sample size decreases. Similarly, when the MDE increases, the sample size decreases. Significance level also affects sample size, but you should not set significance below 95% unless you are very confident in your tolerance for error (see Section 9.)

To find a large effect, you should avoid testing incremental changes, and plan to run more dramatic experiments. Swing for the fences to uncover large changes quickly. Avoid looking for inspiration in case studies that list a small change that garnered a huge effect. Intentionally seek out large changes in your experiment design by testing large changes.

Statistical glossary

CONFIDENCE INTERVAL

Computed interval used to describe the certainty of an estimate of some underlying parameter. In the case of A/B testing, these underlying parameters are conversion rates, or improvement rates. Confidence Intervals have a few theoretical interpretations, most practically be the an interval with a certain probability of containing the true improvement.

EFFECT

The difference in conversion rate between a variation treatment and control in an experiment.

EFFECT SIZE

The amount of difference between the original and variant of a test. This is an input in many sample size calculators used for fixed horizon testing (the "MDE".) In Optimizely, this is "Improvement."

ERROR RATE

The chance of finding a conclusive difference between a control and variation in an A/B test by chance alone OR not finding a difference when there is one. This encompasses both Type I and Type II errors, or false positives and false negatives, respectively.

FALSE POSITIVE RATE

Computed by dividing the number of false positives by (total number of false positives + true negatives.)

FALSE DISCOVERY RATE

The expected number of false discoveries—incorrect winners and losers—computed by dividing the number of false positives by the total number of significant results.

FIXED HORIZON HYPOTHESIS TEST

A hypothesis test designed for an experimenter making a decision at one moment in time (ideally a preset sample size.)

HYPOTHESIS TEST

Sometimes called a t-test, a statistical inference methodology used to determine if an experiment result was likely due to chance alone. Hypothesis tests try to disprove a null hypothesis, the assumption that two variations are the same. In the context of A/B testing, Hypothesis tests will help determine the probability that one variation is better than the other, supposing the variations were actually the same.

IMPROVEMENT

Sometimes known as 'lift', a performance change for an experimental treatment (variation) in either the positive or negative direction. This could mean a 10% increase or decrease in conversion rate (so, a negative improvement).

NULL HYPOTHESIS

The default hypothesis for evaluating the statistical significance of hypothesis test results. Experimenters assume that their experimental treatment (variation) will perform the same as the original. The goals of the hypothesis test is to disprove this null hypothesis that the two variations are the same.

P-VALUE

A value that answers the question: If the null hypothesis for were true, how likely would it be that I witnessed improvement I just saw? In other words, if two variations were indeed the same, how likely is it that the observed conversion rate difference (improvement) was due to random chance? This can also be thought of as the type 1 error rate of a specific test, if the two variations are indeed the same.

SAMPLE SIZE

A method for reducing type I error in hypothesis testing under the assumption of a fixed horizon test. Setting a sample size for a test before starting the experiment sets expectations for how long an experiment should collect data before computing the results.

SAMPLE SIZE CALCULATOR

A method for reducing type I error in hypothesis testing under the assumption of a fixed horizon test. Setting a sample size for a test before starting the experiment sets expectations for how long an experiment should collect data before computing the results.

SEQUENTIAL HYPOTHESIS TEST

Another type of hypothesis testing where an experimenter can make a decision about their test at any time. In this case, there is no "horizon" for the test, and continuous monitoring does not introduce the risk of increased false positives (errors) as it would in a fixed horizon hypothesis test.

STATISTICAL CONFIDENCE

The confidence that a null hypothesis is not true. In other words, it can be thought of as the "chance" that or "confidence" that a variation is different than the variation. It is calculated as (1 - p-value) and in Optimizely, renamed as "Chance to Beat Baseline."

STATISTICAL ERROR

A result that reaches statistical significance that does not represent a significant result. This is effectively the inverse of statistical significance. If an experiment variation reaches 95% statistical significance, for example, there is a 1 in 20 chance that the effect was found because of spurious trends in the experiment data.

STATISTICAL POWER

Sometimes expressed as 1 - type II error, this is the probability that an experimenter will detect a difference when it exists. It is also the probability of correctly rejecting a null hypothesis. In Optimizely's Stats Engine, all experiments are adequately powered.

STATISTICAL SIGNIFICANCE

The number in an experiment results report that represents the likelihood that the difference between a variation and the original (the effect) is not due to chance. This is displayed as a value between 0 and 99%, or 1 minus the false discovery rate. (We'll discuss these concepts in greater detail later.)

STATISTICAL SIGNIFICANCE LEVEL

The threshold of p-values an experimenter will accept. In the case where the p-value threshold is ≤ .05, statistical significance is displayed as 95%. This threshold describes the level of error an experimenter is comfortable with in a given experiment.

TYPE I ERROR

Occurs when a conclusive result (winner or loser) is declared, and the test is actually inconclusive. This is commonly termed a "false positive." Where "positive" is more precisely described as conclusive (can be either a winner or loser.) Hypothesis tests that calculate statistical significance are usually doing so to control these type 1 errors in experiments they run.

TYPE II ERROR

No conclusive result (winner or loser) is declared, failing to discover a conclusive difference between a control and variation when there was one. This is also termed a "false negative."

TRADITIONAL STATISTICS

This is how we'll refer to statistics that were developed to support experimentation in offline situations. These methods are still used by many online A/B testing platforms to calculate statistical significance. They are also called traditional hypothesis testing or fixed-horizon testing.

To learn more about best practices for people, process, and technology for a winning optimization strategy, download a copy of our Roadmap to Building an Optimization Program.

Experience Optimization can be instrumental in improving the performance of your marketing campaigns, product development, and more. Learn how to convert more visitors into customers by reading The Experience Optimization Playbook.

ABOUT THIS GUIDE

A Practical Guide to Statistics

WRITTEN BY
Shana Rusonis
Content Marketing Specialist, Optimizely
@srusonis

DESIGNED BY
Jessie Ren
Communication Designer, Optimizely
@backtofutura

THANK YOU TO
Ural Cebeci, Darwish Gani, Tommy Giglio, Andrew Gori, Ramesh Johari, Robin Pam, Leonid Pekelis, Sean Oliver, David Walsh

TO LEARN MORE ABOUT OPTIMIZELY, SCHEDULE A LIVE DEMO TODAY AT OPTIMIZELY.COM/DEMO

Optimizely

About Optimizely

Optimizely is the world's leading optimization platform, providing A/B testing, multivariate testing, and personalization for websites and iOS applications. The platform's ease of use empowers organizations to conceive of and run experiments that help them make better data-driven decisions. With targeting and segmentation using powerful real-time data, Optimizely meets the diverse needs of any business looking to deliver unique experiences to their visitors.

San Francisco Office
631 Howard Street, Suite 100
San Francisco, CA 94105

Amsterdam Office
Nes 76
1012 KE Amsterdam
The Netherlands