How long to run an experiment

Skip Ahead

- The importance of sample size
- Optimizely’s Sample Size Calculator
  - Baseline conversion rate
  - Minimum detectable effect (MDE)
  - Statistical significance
- The right level of risk for you
- Why isn’t my experiment reaching significance?
- Visitor segments

This article will help you:

- **Improve your planning** by learning how sample size affects experiment duration
- Use Optimizely’s Sample Size Calculator to **calculate the length of time** an experiment will likely run based on sample size
- Choose a **baseline conversion rate** and **minimum detectable effect** for your planning estimates

You have a theory about how to improve your conversion rate, you’ve built your test, and you’re ready to turn it on. Congratulations! So, how long do you have to wait to know if your theory is correct? Traditionally, you had to figure out the total sample size you need, divide it by your daily traffic, then stop the test at the exact sample size that you calculated.

Optimizely’s Stats Engine removes the **requirement** to calculate the sample size you need in advance because it collects evidence as your test runs to declare significant results and show you winners and losers as quickly and accurately as possible.

Even so, you can plan more accurately if you understand how sample size affects experiment length and can estimate experiment length in advance. Read on to learn how, or head to our Optimizely Academy course on planning.
and prioritizing experiments.

The importance of sample size

Even though you no longer need to calculate sample size as an experiment runs, you should understand why it’s important to have a healthy sample size when making decisions.

A healthy sample size is at the heart of making accurate statistical conclusions and a strong motivation behind why we created Stats Engine. When your test has a low conversion rate for a given sample size, it means that there is not yet enough evidence to conclude that the effect you’re seeing is due to a real difference between the baseline and variation instead of chance. In statistical terms, your test is underpowered.

The table below estimates the sample size you would need to accurately detect different levels of Improvement (relative difference in conversion rates) across a few different baseline conversion rates based on Optimizely’s Sample Size Calculator and Stats Engine. It takes fewer visitors to detect large differences in conversion rates—look across any row to see how it works.

The same is true for higher baseline conversion rates: as your baseline conversion rate gets higher, you need a smaller sample size to measure Improvement. Read each column from top to bottom to see how this works.
Stats Engine lets you evaluate results as they come in and avoid making decisions on tests with low, underpowered sample sizes (a "weak conclusion"), without committing to predetermined sample sizes before running a test. You want to avoid making business decisions based on underpowered tests because any improvement that you see is unlikely to hold up when you implement your variation, which could result in spending valuable resources and realizing no benefit.

As you're running experiments, Optimizely shows you an estimate of how many visitors you'll need to reach statistically significant results.

When your variation reaches a statistical significance greater than your desired significance level (by default, 90%), Optimizely will declare the variation a winner or loser. You can stop the test when your variations reach significance.

If some of your variations haven't reached significance, decide whether you can afford to wait for the number of visitors needed to reach significance or use the Sample Size Calculator to calculate how many visitors you would need if the Improvement percentage changes.

You'll see a high Improvement percentage with a Statistical Significance of 0% if your experiment is underpowered and hasn't had enough visitors. As more visitors encounter your variations and convert, you'll start to see Statistical Significance increase because Optimizely is collecting evidence to declare winners and losers.

Even with Stats Engine in place, you probably still want to know how long you can expect your experiments to take for planning. This article will walk you through the process.

**Optimizely's Sample Size Calculator**

Use our [Sample Size Calculator](#) to determine how much traffic you will need for your conversion rate experiments. It's useful for estimating experiment length in advance, which helps with planning. Also, other calculators that account for traditional fixed-horizon testing will not give you an accurate estimate of Optimizely's test duration.

Based on two inputs (baseline conversion rate and minimum detectable effect), the calculator returns the sample sizes you need for your original and your variation to meet your statistical goals. You can also change the statistical significance, which should match the statistical significance level you choose for your Optimizely project. The values you input for the calculator will be unique to each experiment and goal.
Here’s what the calculator looks like. To help you use the calculator for your conversion rate test, we’ll walk you through each input in more detail below.

Great, I’m done calculating sample size! Now, how long will it take to run my experiment?

You’ll translate sample size into the estimated number of days to run your experiment with two calculations:

Calculation #1  Calculation #2

Sample size  Total number of visitors you need
× Number of variations in your experiment  + Average number of visitors per day

Total number of visitors you need  Estimated number of days to run experiment

If you’re trying to calculate experiment length, but your site has low traffic, check out some strategies in Testing tips for low-traffic sites.

Baseline conversion rate

Baseline conversion rate is the current conversion rate for the page you’re testing. Conversion rate is the number of conversions divided by the total number of visitors.

You can usually calculate baseline conversion rates with data from analytics platforms like Google Analytics or from a previous Optimizely experiment. If you don’t have a previous Optimizely experiment, you can run a monitoring campaign: an Optimizely experiment that has only an original, and no variations, to measure baseline conversions.

Minimum detectable effect (MDE)

This is a simple idea, but a long explanation. If you play with the Sample Size Calculator, it will probably become clear pretty quickly, and then you can skip this long explanation.
After you entered your baseline conversion rate in the calculator, you need to decide how much change from the baseline (how big or small a lift) you want to detect. You’ll need less traffic to detect big changes and more traffic to detect small changes. The Optimizely Results page and Sample Size Calculator will measure change relative to the baseline conversion rate.

To demonstrate, let’s use an example with a 20% baseline conversion rate and a 5% MDE. Based on these values, your experiment will be able to detect 80% of the time when a variation's underlying conversion rate is actually 19% or 21% (20% ± 5% × 20%). If you try to detect differences smaller than 5%, your test is considered underpowered.

Power is a measure of how well you can distinguish the difference you are detecting from no difference at all. So running an underpowered test is the equivalent of not being able to strongly declare whether your variations are winning or losing.

Remember, your experiment’s primary metric determines whether a variation “wins” or “loses”—it tracks how your changes affect your visitors’ behaviors. Learn more about primary metrics in Primary and secondary metrics and monitoring goals.

In Optimizely, effect (or lift) is labeled Improvement on the Results page. Effect (or lift) is always presented as relative, not absolute.

If you enter the baseline conversion rate and MDE into the Sample Size Calculator, the calculator will tell you what sample size you need for your original and each variation. The calculator's default setting is the recommended level for statistical significance for your experiment. You can change the statistical significance value according to the right level of risk for your experiment.

You can also use MDE to benchmark how long to run a test and the impact you’re likely to see. This approach can help provide guidelines, in spite of the uncertainty of testing, so you can prioritize experiments according to expected return on investment. To learn more, read "Use MDE to prioritize tests."

**Statistical significance**

Statistical significance answers the question, “How likely is it that my experiment results will say I have a winner when I actually don’t?” We usually consider 90% statistical significance. Another way to say the same thing is that we will accept a 10% false positive rate, where the result is not real (100% - 10% = 90%).

The Sample Size Calculator defaults to 90% statistical significance, which is generally how experiments are run. You can increase or decrease the level of statistical significance for your experiment, depending on the right level of risk for you.

You can change the statistical significance level that Optimizely uses to declare winners and losers for your experiments under Settings > Advanced:
Does Optimizely use 1-tailed or 2-tailed tests?

In A/B testing, a 1-tailed test tells you whether a variation can identify a winner. A 2-tailed test checks for statistical significance in both directions. Previously, Optimizely used 1-tailed tests because we believe in giving you actionable business results, but we now solve this for you even more accurately with false discovery rate control.

The right level of risk for you

When you're running an experiment, you may need to consider the trade-off between running experiments quickly and reducing the chance of inaccuracy in your results (false positives and false negatives). Experiments are usually run at 90% statistical significance. You can adjust this threshold based on how much risk of inaccuracy you can accept.

At the end of the day, you should be aware of the tradeoff between accurate data and available data when making time-sensitive business decisions based on your experiments. For example, imagine your experiment requires a large sample size to reach statistical significance, but you need to make a business decision within the next 2 weeks. Based on your traffic levels, your test may not reach statistical significance within that timeframe. What do you do? If your organization feels that the impact of a false positive (incorrectly calling a winner) is low, you may decide to decrease the statistical significance to see results declared more quickly.

Why isn’t my experiment reaching significance?

In general, smaller differences take longer to detect because you need more data to confirm that Optimizely observed an actual, statistically significant difference rather than random changes in conversion patterns.

If your experiment has been running for a considerable amount of time and you still need more unique visitors to reach significance, this could be because Optimizely is observing scattered data—conversions that are erratic and inconsistent over time. If your data has high variability, Stats Engine will require more data before showing significance.

When you are measuring impulse-driven goals like video plays or e-mail sign-ups, data tends to be more scattered because visitor behavior tends to be erratic and easily affected by many small impulses. However, when you are
measuring goals that involve carefully weighed decisions, such as a high-value purchase, you will see more stable, less variable data. Optimizely’s Stats Engine automatically calculates variability and adjusts accordingly.

Here’s an example of data variability:

![Graph showing Low and High Variability Data]

**Low Variability Data:** The blue line shows a data set for which the baseline conversion rate varies from 3.2% to 4.8%. If a variation raises this metric to 5%, we can tell that it is significant.

**High Variability Data:** The green line shows data set whose baseline conversion rate varies between 2% and 6%. If a variation raises this metric to 5%, we will need more data to call results significant because 5% falls within the baseline conversion range.

**Visitor segments**

As we mentioned, not all visitors behave like your average visitors, and visitor behavior can affect statistical significance. For example, an experiment that tests a pop-up promotional offer may generate positive lift **overall**, but be a statistically significant loss among visitors on mobile devices because the pop-up is difficult to close on small screens.

Optimizely lets you filter your results so you can see if certain groups of visitors behave differently from your visitors overall. This is called **segmenting**. With segmenting, you can discover insights that will help you run more effective experiments. To continue our example, when you run similar experiments on pop-up promotions in the future, you might exclude mobile visitors based on what you learned.