Ensuring efficiency and accuracy in signup form A/B testing

Author: Michael Lawson

Claps: 462

Date: Oct 16, 2020

In many ways, A/B tests are central to the task of marketing. Have multiple ideas for your creative approach but youâ€™re not sure which one will resonate with your audience best? Try them both out and see which performs better. Curious whether a discount will lead to enough extra signups to justify the cost? Add one, but protect your brand and make sure only a portion of the audience gets to see the discount. A/B testing provides a data-driven method for optimizing your content and improving your marketing funnel.

While A/B tests may be simple in principle, not all A/B tests are identical. An A/B test for changing the font on your entire homepage will work differently from an A/B test on the subject lines in an email campaign. Every distinct flavor of A/B testing will have its own peculiarities.

With that in mind, here are a few of the big statistical concepts you should think about when designing signup form A/B testing.

Quantifying evidence

One of the biggest issues in an A/B test â€” arguably the single biggest question to answer â€” is how much evidence we have that a variation is performing better than all the other variations. In form A/B testing, what that translates to is: how sure are we that the submit rate for a variation is truly the highest submit rate among all the variations in the test?

There are a number of ways to quantify evidence, but one of the most helpful is the *Bayesian win probability*. We could spend the entire rest of this blog walking through the ways that itâ€™s different from a frequentist p-value. This is what you likely learned about when your teacher mentioned â€œp-valuesâ€� in statistics courses in school. Suffice it to say: the win probability gives us a direct estimate of what we want. Bayesian win probability tells us, under reasonable and minimally-informative assumptions, how likely it is given our data that the current leading variation truly has the highest submit rate, after accounting for random chance.

Win probabilities are probabilities, which means that they must lie somewhere between 0 and 100%. The floor for a win probability is 1/K, where K is the number of variations in your test. For instance, in a test with 2 variations, a win probability of 50% is saying that â€œeven with the data youâ€™ve gathered, itâ€™s equivalent to a coin flip.â€�

If the win probability is high, it represents strong evidence that the variation that is currently ahead is truly better. We use the threshold of 90% win probability to indicate evidence strong enough that you may want to stop your test, which we call *statistical significance*.Â[1]

Measuring precision

Win probability by itself doesnâ€™t tell us the whole story. Letâ€™s think of a simple example: suppose variation A has 6 views and 0 submits, and variation B has 6 views and 6 submits. If we look at the win probability alone, it will be greater than 99%, so we might be tempted to end the test now.

Of course, in practice, thatâ€™s not a decision weâ€™d want to make â€” weâ€™d be basing our decision off a total of 12 views! This result could easily be due to an anomaly, random chance, or similar factors â€” we want to give ourselves more of a chance to assess the real trend and get results we can trust.

Simply put, weird things can happen early in a test. To protect from early-test weirdness, we take the following two steps.

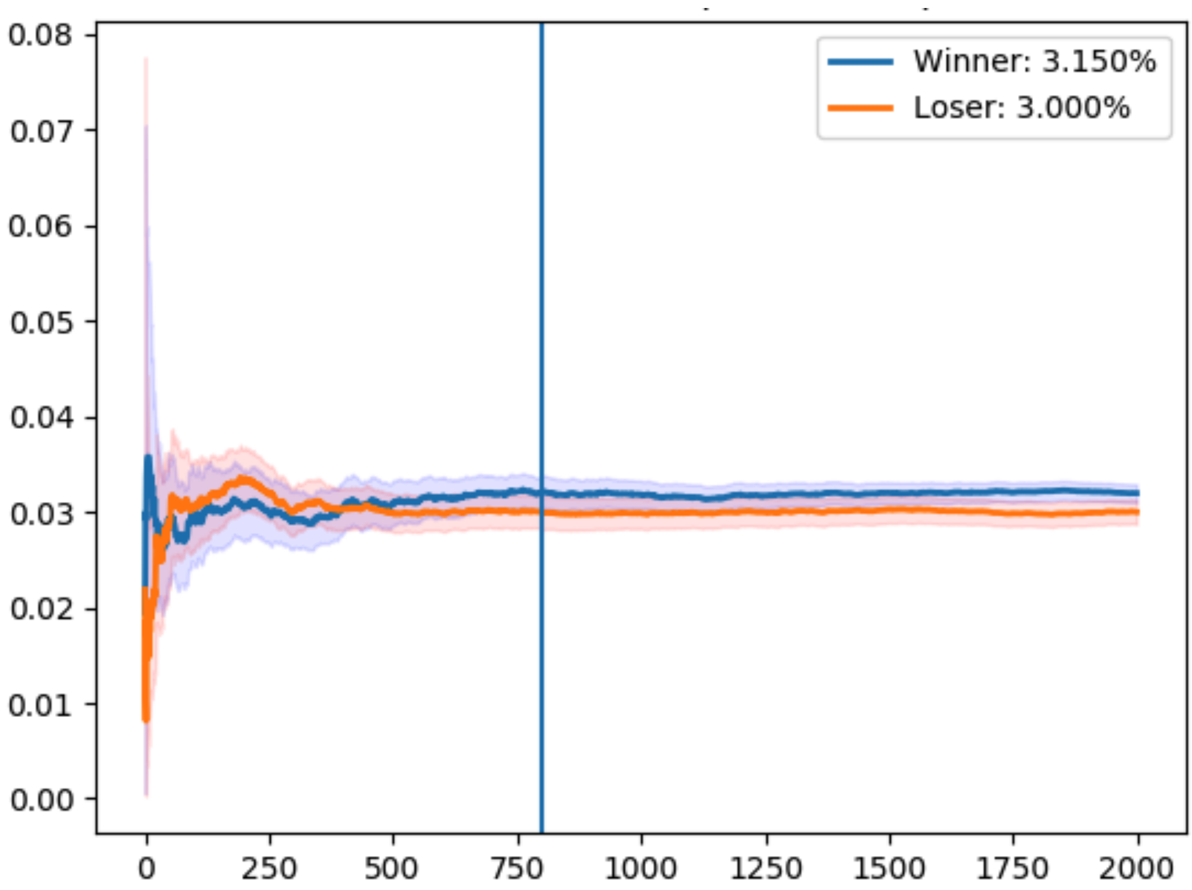
First, we make sure a sufficient level of data accrues during the testing phase. We ran a suite of simulations based on realistic scenarios for customers of all sizes. What we found was that requiring 2000 views in every variation prevented us from making decisions too early, but still allowed tests to end in a reasonable time frame.

You might ask: *Why run simulations? Arenâ€™t there pre-set best practices for this type of test?* While there are best practices, theyâ€™re meant to apply across all sorts of tests and every type of data structure. The signup forms data structure isnâ€™t just any data structure, and signup form A/B tests arenâ€™t just any type of test â€” we can use our knowledge about the specific details of signup form testing to make our tests more efficient while still maintaining experimental rigor.

Second, we consider the precision of our estimates regarding the submit rate of each variation. To do this, we look at the 95% posterior credible intervals of form submit rate for each variation. Credible intervals are essentially the Bayesian version of a 95% confidence interval. We check how much these credible intervals overlap between the leading variation and the next-best variation to arrive at our conclusion.

To give some intuition: we are essentially saying that we want our estimates to be precise enough that we really believe form submit rate A is higher than form submit rate B. If the credible intervals â€” the range of values we believe the form submit rate could reasonably be â€” overlap by a lot, then it is hard for us to say the form submit rates are actually different. But if they donâ€™t overlap much, or at all, then it is highly unlikely that the form submit rates are actually the same.

For example, letâ€™s consider the following graph of cumulative form submit rate by hours the test has run in one of our simulations. This is exactly the sort of graph you would see in your A/B test results page today, but with the 95% credible intervals added to the line for each variation. Since this is a simulation, we know in this test that the blue line has the higher true submit rate. So, logically, the blue line should win. Early on, until around hour 400, though, the orange line appears to be leading! But take a closer look at those credible intervals â€” the amount of overlap is huge. That tells us that, even though the orange line *looks* ahead, we donâ€™t have enough trust in our estimates yet to actually *know that it is* ahead. Sure enough, as more data comes in and our estimates get more precise, we find the actual winner. The vertical line shows the point at which the credible intervals overlap by a small enough amount for us to be confident in our result.



One simulated form A/B test, with form submit rates for each variation and their 95% credible intervals shown for each hour the test ran. In this test, the true winning variation has a submit rate of 3.15% and the true trailing variation has a submit rate of 3.0%. The losing variation outperforms the true winner at first, but not significantly â€” note the large overlap between the credible intervals. The intervals only separate to 10% overlap at the vertical line, at which point the true winner has taken over.

We want our estimates to be precise, so we look for the overlap between the credible intervals to be small â€” 10% of the width of the narrower of the two intervals. That number is also based on our simulations â€” it provides the best tradeoff between the false positive rate and the expected length that the test would run.

Finding significance

Naturally, we donâ€™t want to overload our reporting screens with lots of numbers that you have to track down and weigh at the same time. Weâ€™ve condensed our criteria above into a single measure of statistical significance. A test has reached statistical significance if all of the following are true:

- Win probability >= 90%
- 95% credible intervals for the best and second-best variations overlap <= 10%
- Minimum number of views in any variation >= 2000

When a test reaches statistical significance, you should be confident that you have learned all you can, and can therefore move on to the next test. If you agree with our logic, you can take a shortcut and set your tests up to end automatically as soon as they have reached statistical significance â€” saving you time while still providing results you can trust.

Maximize while you run

Thatâ€™s not the end of the story, though. We have already discussed ways to know when to end your test â€” but what about maximizing performance *while* your test is running?

Letâ€™s consider an extreme case to understand how important this could be. Letâ€™s say you have three variations of the main signup form on your homepage: one with a red button, one with a blue button, and one with a black button. However, you also have a black background image, which makes the black button *really* hard to spot. In a traditional A/B test, you would set up those variations to run with 1/3 of the traffic each, and you would keep them that way as long as it takes to find results. Your red button and blue button could both be going strong with 5% submit rates, but your black button might be lagging behind at 1%. This is, of course, because it is so much harder to figure out how to submit the form with a black button on a black background. Every day that you keep the black button at 33% of your website traffic , you lose submits, email addresses, and sales.

Thankfully, thereâ€™s a better way. Suppose you let the test gather data for a few days, then notice that the black button is performing substantially worse than the others. At that point, you can shift the allocations away from the black button. Leave some, because as we talked about, weird things can happen at the start of a test. We also donâ€™t want to give a slow-starting variation no chance to catch up. But maybe only 10% of traffic sees that form, and the 23% you just freed up gets divided between the red button and the blue button. Not only are you now capturing more signups by having more of your traffic see the variations that perform better, youâ€™re also boosting your speed to find a winner between the red and blue button by making the test larger. Itâ€™s a win-win, and you only have to make one simple change.

If we make that approach a bit smarter and more automatic, we end up with what is known as the *multi-armed bandit* approach. We use the multi-armed bandit approach to maximize signups while the test is running and drive traffic to understand the differences between the variations we care about the most. Best of all, we do it without sacrificing performance.

Since weâ€™re already discussing Bayesian win probabilities, we should mention that we use a Bayesian multi-armed bandit method called Thompson Sampling to decide how traffic will be divided in our form A/B tests. Without diving too far into the math, the important thing to note is that this method is best-in-class for multi-armed bandit methods: there is a theoretical mathematical upper bound on performance that multi-armed bandit algorithms can reach, and Thompson Sampling reaches it.

Putting it all together

All A/B tests require something special, and all of them offer unique opportunities. For form A/B tests in particular, you can:

- Make them more automatic and less manual, by ending automatically after a sufficient level of confidence is reached.
- Make them smarter, by making data-driven decisions at the first moment youâ€™re ready to do so.
- Make them more valuable, gathering more signups while you test.

All together, you can optimize your performance while you optimize your form, all while reducing manual work. It just takes a bit of data science magic!

Whatâ€™s next?

As we said at the start: not all A/B tests are identical. In our minds, thatâ€™s actually a strength of the A/B testing framework: itâ€™s not meant to be one-size-fits-all. A/B testing should be adapted to the circumstances it lives in, allowing you to glean accurate, scientifically-driven insights all throughout your marketing funnel.

Our plans for using A/B testing effectively are not limited to signup forms. Other parts of the marketing funnel, from one-shot campaigns to automated emails, can gain similar efficiencies and quality-of-life improvements with the right approach. Experimentation â€” trying things out, gathering data, and drawing intelligent conclusions â€” is core to marketing. We wonâ€™t be satisfied until the ideal experience for running an experiment is available to marketers everywhere they need it.