Harvard
Business
Review

Analytics And Data Science

A Refresher on A/B Testing

by Amy Gallo

June 28, 2017

Summary. A/B testing is a way to compare two versions of something to figure out which performs better. While it's most often associated with websites and apps, the method is almost 100 years old and it's one of the simplest forms of a randomized controlled experiment.... **more**

It's all about data these days. Leaders don't want to make decisions unless they have evidence. That's a good thing, of course, and fortunately there are lots of ways to get information without having to rely on one's instincts. One of the most common methods, particularly in online settings, is A/B testing.

To better understand what A/B testing is, where it originated, and how to use it, I spoke with Kaiser Fung, who founded the applied analytics program at Columbia University and is author of Junk Charts, a blog devoted to the critical examination of data and graphics in the mass media. His latest book is *Number Sense: How to Use Big Data to Your Advantage.*

What Is A/B Testing?

A/B testing, at its most basic, is a way to compare two versions of something to figure out which performs better. While it's most often associated with websites and apps, Fung says the method is almost 100 years old.

In the 1920s statistician and biologist Ronald Fisher discovered the most important principles behind A/B testing and randomized controlled experiments in general. "He wasn't the first to run an experiment like this, but he was the first to figure out the basic principles and mathematics and make them a science," Fung says.

Fisher ran agricultural experiments, asking questions such as, What happens if I put more fertilizer on this land? The principles persisted and in the early 1950s scientists started running clinical trials in medicine. In the 1960s and 1970s the concept was adapted by marketers to evaluate direct response campaigns (e.g., would a postcard or a letter to target customers result in more sales?).

A/B testing, in its current form, came into existence in the 1990s. Fung says that throughout the past century the math behind the tests hasn't changed. "It's the same core concepts, but now you're doing it online, in a real-time environment, and on a different scale in terms of number of participants and number of experiments."

How Does A/B Testing Work?

You start an A/B test by deciding what it is you want to test. Fung gives a simple example: the size of the subscribe button on your website. Then you need to know how you want to evaluate its performance. In this case, let's say your metric is the number of visitors who click on the button. To run the test, you show two sets of users (assigned at random when they visit the site) the different versions (where the only thing different is the size of the button) and determine which influenced your success metric the most. In this case, which button size caused more visitors to click?

In real life there are lots of things that influence whether someone clicks. For example, it may be that those on a mobile device are more likely to click on a certain size button, while those on desktop are drawn to a different size. This is where randomization can help — and is critical. By randomizing which users are in which group, you minimize the chances that other factors, like mobile versus desktop, will drive your results on average.

"The A/B test can be considered the most basic kind of randomized controlled experiment," Fung says. "In its simplest form, there are two treatments and one acts as the control for the other." As with all randomized controlled experiments, you must estimate the sample size you need to achieve a statistical significance, which will help you make sure the result you're seeing "isn't just because of background noise," Fung says.

Sometimes, you know that certain variables, usually those that are not easily manipulated, have a strong effect on the success metric. For example, maybe mobile users of your website tend to click less on anything, compared with desktop users. Randomization may result in set A containing slightly more mobile users than set B, which may cause set A to have a lower click rate regardless of the button size they're seeing. To level the

playing field, the test analyst should first divide the users by mobile and desktop and then randomly assign them to each version. This is called blocking.

The size of the subscribe button is a very basic example, Fung says. In actuality, you might not be testing just the size but also the color, and the text, and the typeface, and the font size. Lots of managers run sequential tests — e.g., testing size first (large versus small), then testing color (blue versus red), then testing typeface (Times versus Arial) — because they believe they shouldn't vary two or more factors at the same time. But according to Fung, that view has been debunked by statisticians. And sequential tests are suboptimal because you're not measuring what happens when factors interact. For example, it may be that users prefer blue on average but prefer red when it's combined with Arial. This kind of result is regularly missed in sequential A/B testing because the typeface test is run on blue buttons that have "won" the prior test.

Instead, Fung says, you should run more-complex tests. This can be hard for some managers, since the appeal of A/B tests are how straightforward and simple they are to run (and many people designing these experiments, Fung points out, don't have a statistics background). "With A/B testing, we tend to want to run a large number of simultaneous, independent tests," he says, in large part because the mind reels at the number of possible combinations you can test. But using mathematics you can "smartly pick and run only certain subsets of those treatments; then you can infer the rest from the data." This is called "multivariate" testing in the A/B testing world and often means you end up doing an A/B/C test or even an A/B/C/D test. In the example above with colors and size, it might mean showing different groups: a large red button, a small red button, a large blue button, and a small blue button. If you wanted to test fonts, too, the number of test groups would grow even more.

How Do You Interpret the Results of an A/B Test?

Chances are that your company will use software that handles the calculations, and it may even employ a statistician who can interpret those results for you. But it's helpful to have a basic understanding of how to make sense of the output and decide whether to move forward with the test variation (the new button in the example above).

Fung says that most software programs report two conversion rates for A/B testing: one for users who saw the control version, and the other for users who saw the test version. "The conversion rate may measure clicks, or other actions taken by users," he says. The report might look like this: "Control: 15% (+/- 2.1%) Variation 18% (+/- 2.3%)." This means that 18% of your users clicked through on the new variation (perhaps your larger blue button) with a margin of error of 2.3%. You might be tempted to interpret this as the actual conversion rate falling between 15.7% and 20.3%, but that wouldn't be technically correct. "The real interpretation is that if you ran your A/B test multiple times, 95% of the ranges will capture the true conversion rate — in other words, the conversion rate falls outside the margin of error 5% of the time (or whatever level of statistical significance you've set)," Fung explains.

If this is hard to wrap your head around, join the club. What's important to know is that the 18% conversion rate isn't a guarantee. This is where your judgment comes in. An 18% conversation rate is certainly better than a 15% one, even allowing for the margin of error (12.9%–17.1% versus 15.7%–20.3%). You might hear people talk about this as a "3% lift" (lift is simply the percentage difference in conversion rate between your control version and a successful test treatment). In this case, it's most likely a good decision to switch to your new version, but that will depend on the costs of implementing the new version. If they're low, you might try out the switch and see what happens in

actuality (as opposed to in tests). One of the big advantages to testing in the online world is that you can usually revert back to your original pretty easily.

How Do Companies Use A/B Testing?

Fung says that the popularity of the methodology has risen as companies have realized that the online environment is well suited to help managers, especially marketers, answer questions like, "What is most likely to make people click? Or buy our product? Or register with our site?" A/B testing is now used to evaluate everything from website design to online offers to headlines to product descriptions. (In fact, last week I looked at the results of A/B testing on the language we use to market a new product here at HBR.)

Most of these experiments run without the subjects even knowing. "As a user, we're part of these tests all the time and don't know it," Fung says.

And it's not just websites. You can test marketing emails or ads as well. For example, you might send two versions of an email to your customer list (randomizing the list first, of course) and figure out which one generates more sales. Then you can just send out the winning version next time. Or you might test two versions of ad copy and see which one converts visitors more often. Then you know to spend more getting the most successful one out there.

What Mistakes Do People Make When Doing A/B Tests?

I asked Fung about the mistakes he sees companies make when performing A/B tests, and he pointed to three common ones.

First, he says, too many managers don't let the tests run their course. Because most of the software for running these tests lets you watch results in real time, managers want to make decisions

too quickly. This mistake, he says, “evolves out of impatience,” and many software vendors have played into this overeagerness by offering a type of A/B testing called “real-time optimization,” in which you can use algorithms to make adjustments as results come in. The problem is that, because of randomization, it’s possible that if you let the test run to its natural end, you might get a different result.

The second mistake is looking at too many metrics. “I cringe every time I see software that tries to please everyone by giving you a panel of hundreds of metrics,” he says. The problem is that if you’re looking at such a large number of metrics at the same time, you’re at risk of making what statisticians call “spurious correlations.” In proper test design, “you should decide on the metrics you’re going to look at before you execute an experiment and select a few. The more you’re measuring, the more likely that you’re going to see random fluctuations.” With so many metrics, instead of asking yourself, “What’s happening with this variable?” you’re asking, “What interesting (and potentially insignificant) changes am I seeing?”

Lastly, Fung says that few companies do enough retesting. “We tend to test it once and then we believe it. But even with a statistically significant result, there’s a quite large probability of false positive error. Unless you retest once in a while, you don’t rule out the possibility of being wrong.” False positives can occur for several reasons. For example, even though there may be little chance that any given A/B result is driven by random chance, if you do lots of A/B tests, the chances that at least one of your results is wrong grows rapidly.

This can be particularly difficult to do because it is likely that managers would end up with contradictory results, and no one wants to discover that they’ve undermined previous findings, especially in the online world, where managers want to make changes — and capture value — quickly. But this focus on value

can be misguided, Fung says: “People are not very vigilant about the practical value of the findings. They want to believe that every little amount of improvement is valuable even when the test results are not fully reliable. In fact, the smaller the improvement, the less reliable the results.”

It’s clear that A/B testing is not a panacea. There are more complex kinds of experiments that are more efficient and will give you more reliable data, Fung says. But A/B testing is a great way to gain a quick understanding of a question you have. And “the good news about the A/B testing world is that everything happens so quickly, so if you run it and it doesn’t work, you can try something else. You can always flip back to the old tactic.”

Amy Gallo is a contributing editor at Harvard Business Review, cohost of the *Women at Work* podcast, and the author of two books: *Getting Along: How to Work with Anyone (Even Difficult People)* and the *HBR Guide to Dealing with Conflict*. She writes and speaks about workplace dynamics. Watch her TEDx talk on conflict and follow her on LinkedIn.

𝕏 @amyegallo

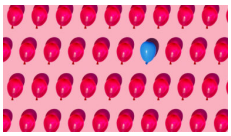
Recommended For You

PODCAST

Employee Performance vs. Company Values: A Manager's Dilemma

A Refresher on Randomized Controlled Experiments

A Refresher on Statistical Significance



An Introduction to Data-Driven Decisions for Managers Who Don't Like Math