p-Hacking your A/B tests

Faking Results · Faking A/B Tests · p-Hackers · Fooling yourself · Theories

credit [1]

"Fine, just stand there being ominously silent. It doesn't change the fact that I have the lab until 9:00."

FAKING RESULTS USING REAL EXPERIMENTS

You are a pharmaceutical company who has blown through a billion dollars of research and development, only to produce a drug that doesn't work. But you want to get it approved by the FDA* anyway so you can recoup that cost, hapless consumers be damned.**

The FDA requires experimental evidence of the drug's efficacy—a controlled study showing positive results at the 90% statistical confidence level. You need to publish a believable report, even though there isn't a statistically-significant effect. How can you achieve your nefarious goal?

Easy: Run studies repeatedly, until you get a false-positive result. Then publish only that result.

All experiments sometimes give false-positives. An experiment that gives the correct result 95% of the time, still gives the wrong answer 5% of the time. If you run it 15-25 times, you will almost surely hit one of those false results.

This is in fact what pharmaceutical companies used to do to get drugs approved by the FDA. To prevent this behavior, the FDA now requires companies to pre-register their studies and publish all results.***

---

* The United States government agency that approves drugs for sale.

** There are many such examples in the US, such as the decongestant Phenylephrine, with nearly $2B in annual sales, which an FDA advisory panel unanimously concluded[2] is ineffective.[3]

This is also what happens in the social sciences, resulting in The Replication Crisis.[5] An experiment is run once, often with a small number of college students. Occasionally something "interesting" happened, and a journal publishes it. Journals don't wait for other teams to reproduce the result. Nobel Laureates have admitted[6] that this fallacy has debunked their own work, and call for "replication rings" to solve the problem. But scientists are people too, and often prefer the fame from an amazing result to the pain of discovering that they are infamous for perpetuating a false-positive.

Before you shake your finger at them, shake that finger at yourself.

Because you're doing this too.

FAKING YOUR A/B TESTS (UNINTENTIONALLY)

To see how this unfolds, consider this simple example: You're testing whether a coin is fair, i.e. that it comes up heads equally often as tails. Your experiment is to flip it 270 times, measuring how often it comes up heads. Of course we don't expect *exactly* 50% heads, because there's randomness. What is a reasonable range to expect from 270 flips, assuming the coin really is fair? According to the binomial distribution, 90% of the time the result will be between 45% and 55% heads, if the coin is fair.

So, you run the experiment, and you get 57% heads. You conclude the coin is biased, and you say "I'm 90% sure of that." Is this the right conclusion? Probably? Maybe?

*** Although there are plenty of recent cases[4] where companies were allowed to submit a subset of trials, yet were approved anyway, and—big surprise!—were later found to be ineffective.

Now imagine you have 10 coins, and you want to test all of them for fairness. So you run the above experiment, once per coin. 9 of the tests result in "fair coin", but one test shows "biased coin".

Should we conclude that the one coin is biased? Almost surely not. Because even if all coins were fair, we know that 10% of the time the test will incorrectly conclude "biased". So this result of `9 / 1` is exactly the result we'd expect if all coins were fair.

But wait a second… what if in fact 9 coins were fair but 1 were biased? Then this is *also* the most likely result! It could have also come up `8 / 2`, but the `9 / 1` result is the most likely.

So: `9 / 1` is the most likely result *both* if all 10 coins are fair *and* if only 9 coins are fair.

So… what exactly can you conclude from the `9 / 1` result? Nothing, yet, not with confidence. What you conclude is that this procedure is insufficient, and that we need to augment the procedure to correct the issue.

The insight is: **You are making exactly this mistake with your A/B tests.**

You are running a bunch of A/B tests. You're looking for (something like) "90% confidence". Mostly the tests have a negative result. Occasionally one works, maybe one out of ten. And you conclude that was a successful test. But this is exactly what we just did with coin-flipping.

In the real world it's often even worse, like using confidence of 85% or 80% and therefore false-positives are much more likely.

Or you don't even pick a confidence level. You don't decide how much N you need to make a conclusion. Instead you "run the test until we get a result that looks conclusive." This is a new type of mistake.

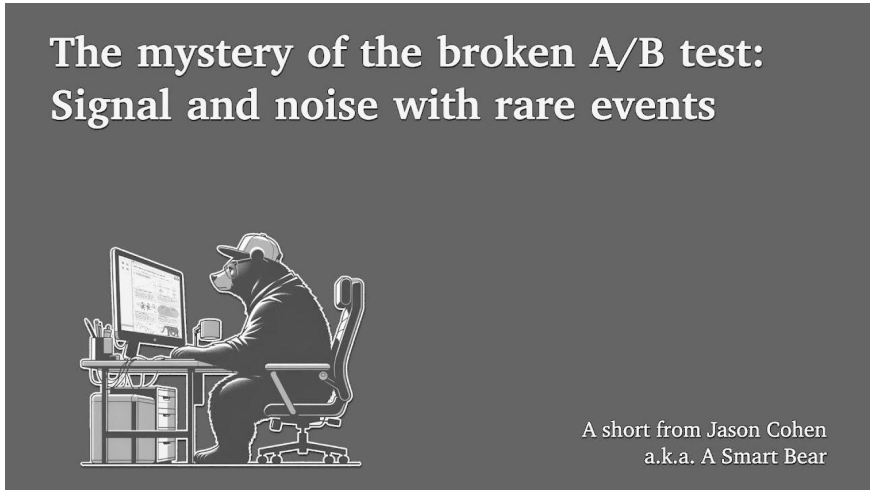

Figure 1: Watch on YouTube[7]

This particular error of "stopping whenever it looks conclusive" is called "p-hacking" by statisticians, and it's been a well-documented fallacy since the 1950s. The reason it's a fallacy, is that when N is small, random fluctuations will often cause a result that looks like "90% confidence of a positive result," whereas if you continue the experiment, the data shifts back into the territory of "negative result."

I show some fun real-world examples of p-hacking, and counter-examples when the experiment is done properly, in this video (Figure 1).

MARKETERS: THE ACCIDENTAL P-HACKERS

Marketers have been making these p-hacking errors in A/B testing for many years. You are too.

We have data. A study* of more than 2100 real-world A/B tests across Optimizely's customer base found a 40% false positive rate. Marketers never knew it; indeed, the Optimizely software declared the tests "significant"! The Marketers never had a chance.

So, roughly half the time you think "this A/B test was successful," it wasn't.

This explains another phenomenon that you're probably familiar with if you've done a lot of A/B testing:

1. You run tests. Sometimes one is significant. You keep that result and continue testing new variants.
2. You keep repeating this process, keeping the designs that are "better."
3. Over time… one is 10% better. Another is 20% better. Another is 10% better.
4. So, that should be 45% better overall.
5. You look back between now and months ago when you first started all this… and you don't see a 45% improvement! Often, there's no improvement at all.

Why didn't all those improvements add up? Because they were false-positives.

HOW TO STOP FOOLING YOURSELF

The easiest thing is to run the test again.

If false positives are 1 in 10 at 90% confidence, then you should be able to run a second test, and get the same result.

---

* Here is the study,[8] and here is a blog post[9] from the author, addressing concerns and caveats.

And don't stop tests early. I know you're excited. Just wait. No p-hacking.

And seek large effects, like double-digit changes in conversion rates. Large effects are unlikely to be caused by randomness; small fluctuations are exponentially more likely to be false-positives. And anyway, large effects actually have an impact on the business, whereas small effects don't. This might mean testing drastic changes instead of incremental ones.

That's it? Almost—**there's a much smarter way to go about this**. And if you want to keep your job even with the rise of AI, you need to be smarter than just running a bunch of variants.

credit[10]

"You won't believe it! Anderson's experiment... the results... they're conclusive! *Conclusive* I tell you!"

FORM A THEORY, TEST THE THEORY, EXTEND THE THEORY

Too often A/B tests are just "throwing shit at the wall." We excuse this behavior by saying "No one knows which headline will work; it's impossible to predict, so we just try things."

Not only is this thoughtless and lazy, it also means **you haven't learned anything, regardless of the result of the test.**

You don't want to be a mindless slinger of random phrases. AI can do that too, and AI isn't a good marketer. Instead, you want to **create validated learning.**

To do this, form a theory, then design experiments to test the theory. Example theories:

1. At this point on the website, visitors are ready to buy, so send them down a purchase funnel with a restricted UX.
2. Here people want to learn more, so talk about options and let them explore features rather than being crammed down a funnel.
3. People are on the fence, so we should be more forceful and confident in our language.
4. People can't see well, especially on mobile devices, so we should have higher contrast colors and less text.
5. Pictures work better than paragraphs, especially since people hate reading and half of them don't speak English natively.
6. People are more likely to click buttons than to click links.
7. People from marketing channel X are more likely to be in a Y state of mind, and to be excited by Z.

Perhaps some theories already popped into your mind. Good! Write those down. Then make designs that would perform better if that theory were true.

It's not "shit on the wall" because this time you have a specific Theory of Customer that your wall-shit is designed to test. And that makes all the difference:

The negative result

Let's say you pick a theory, run a test, and it fails. Is your theory disproved?

Not quite yet. Perhaps your implementation wasn't the best manifestation of the theory. Not extreme enough, or had other issues that covered up the good effects. If you feel this might be the case, run a new experiment.

But if you're still not getting positive results after a few iterations, you have accumulated evidence that the theory is incorrect. That is called "learning." Which wasn't happening when you "threw shit at the wall." Now you know that you need to invent a new, different theory and test *that*.

How useful, and directed.

The positive result

Suppose you had a positive result. Hooray!

Is the theory proven? No, because you read the first half of the article, so you know that positive A/B tests are often false. So, what do you do?

You lean even further into the theory. Run another test that's even more extreme, or a different form of the same concept.

If the theory is truly correct, that will work again, perhaps even better! If it reverts to nothing, you know it wasn't a real result.

Now you're not fooling yourself. You're finding theories that actually correct. That's what "validated learning" looks like.

Since you've actually learned something, you can extend the theory. What else is probably true? What new designs and text and pictures would

leverage those insights even more? Now you might make multiple leaps of improvement, rather than spraying random things on your website.

And you're a smart marketer that AI cannot replace.

Most theories won't be right (or at least not impactful enough to matter). Most tests will come up negative. That's frustrating but also the truth.

You *do* want the truth…[11]

Don't you?

Many thanks to Einar Vollset[12] *for reviewing early drafts.*

Current version of this article:
`https://longform.asmartbear.com/p-hacking/`

More articles & socials:
`https://asmartbear.com`

© 2024 A Smart Bear Press

REFERENCES

1. https://andertoons.com/science/cartoon/8676/fine-stand-there-silent-have-the-lab-until-9
2. https://www.nbcnews.com/health/health-news/fda-panel-says-common-counter-decongestant-phneylephrine-doesnt-work-rcna104424
3. https://www.theatlantic.com/health/archive/2023/09/cold-medicine-decongestant-phenylephrine-ineffective/675303/
4. https://www.biospace.com/6-drugs-approved-despite-failed-trials-or-minimal-data
5. https://en.wikipedia.org/wiki/Replication_crisis
6. https://www.nature.com/articles/nature.2012.11535
7. https://youtu.be/FaUO2-AQmr0
8. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3204791
9. https://www.ron-berman.com/2018/12/17/when-a-paper-becomes-a-sensation/
10. https://andertoons.com/result/cartoon/6798/you-wont-believe-it-experiment-the-results-theyre-conclusive
11. https://longform.asmartbear.com/failure-to-face-the-truth/
12. https://x.com/einarvollset