https://bit.ly/raiseTheBarOnSharedABTests

Raise the Bar on Shared A/B Tests: Make them trustworthy

Ron Kohavi

Apr 13, 2022

If an effect is reliable, any competent researcher should be able to obtain
it when using the same procedures with adequate statistical power
-- Daniel Simons (2014)

It ain't what you don't know that gets you into trouble.
It's what you know for sure that just ain't so
— Anonymous (commonly attributed to Mark Twain)

The major difference between a thing that might go wrong and
a thing that cannot possibly go wrong is that
when a thing that cannot possibly go wrong goes wrong
it usually turns out to be impossible to get at or repair
— The Hitchhiker's Guide to the Galaxy by Douglas Adams

Executive Summary

There is a replication crisis in several scientific fields, with strong evidence that most published research findings are false (Ioannidis 2005, Open Science Collaboration 2015). In software, organizations run thousands of A/B tests every year, with a total of over 100,000 experiment treatments estimated to have been run in 2018 by several of the larger companies (Gupta, et al. 2019).

Sharing A/B test results and identifying common patterns is highly encouraged, as it helps disseminate successful ideas. However, many results being shared are not trustworthy and we must raise the bar on shared A/B tests.

In shared results, we sometimes see little data to help us assess the reliability of experiments. Commercial companies are justified in avoiding the publication of critical numbers, such as conversion rates, which may be material to the business, creating legal and financial risks. That said, publishing only relative improvements without any support for their validity is insufficient for establishing reasonable trust in the results. I propose that for an A/B test to be shared, it must meet the following conditions:

1. The p-value must be below 0.01 for a two-tailed test, which may be achieved through replication.
2. The pre-experiment statistical power parameters should be provided: the power should be at least 80% for a relative treatment effect delta of at most 10% (ideally under 5%).
3. Clear specification of the OEC (Overall Evaluation Criterion).
4. All differences between the treatment(s) and control must be clearly stated.
5. Guardrails evaluated: in particular, at least a test for SRM (Sample-Ratio-Mismatch) must be done. Ideally, more metrics that ensure basic assumptions were not violated.

Introduction

We see amazing tests being published regularly. Here are three I've seen in the last few weeks:

1. Treatment that has no coupon code and uses a longer form to get a massive 140% lift to form fill completion with 90% confidence (3/22/2022).
2. Treatment added a value proposition box, and got a whopping 153% lift in conversions with 95% confidence (3/24/2022).
3. Treatment changed the button text from "Order now" to "Next step" and got a whopping 196% increase to conversions with 99% confidence (3/10/2022).

Jonny Longden in a recent LinkedIn post noted that he has only once seen a test that was trustworthy and doubled conversion rate. He noted that most winning experiments might demonstrate uplifts of perhaps 10%. In another recent LinkedIn post, the author Deborah O'Malley shared the criticism that big lifts are often B.S. (this is also why I'm comfortable sharing some of her examples above, as she's clearly recognizing the limitations of those studies).

So how can we separate the wheat from the chaff? How do we know if to trust a result? The two key criteria that are often missed are low p-value and power.

Low P-Value

P-value is often misinterpreted as the probability of a false positive. I won't dwell on this point, and I will point to an earlier LinkedIn post about this. The key point is that the false positive risk (FPR) depends on the prior probability that your idea will be successful. Based on reported numbers in industry, that prior probability is around 8%-33%. If your rate is 15%, then the probability of a false positive for a properly powered experiment is 15%, not 5%. By requiring published results to pass the bar of p-value <= 0.01 for a two-tailed test, or 0.005 for a one-tailed test, we will significantly reduce the rate of published false positives. This also aligns with the well-known 72-author paper proposing redefining statistical significance (Benjamin, et al. 2017) threshold to 0.005 for "claims of new discoveries." A key point to remember is that the more surprising the result, the more evidence we want in the form of a lower p-value to override the prior that the result is unlikely (Twyman's law). Commercial software typically uses the term "confidence" for 1-pvalue. Of the three examples above, only the 3rd has a confidence of 99%, thus (borderline) passing this criterion.

One way to achieve a lower p-value is to do a replication run. If you have a borderline p-value of 0.05, run it again. If you get another 0.05, the combined value will be below 0.01. A simple approximation for combining two p-values is simply: $p1 * p2 * 2$ (so 0.05*0.05*2=0.005). My spreadsheet for meta-analysis is available here.

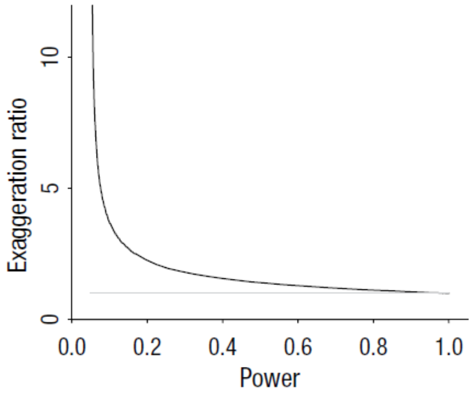
Statistical Power

This is the most critical part that is often missed in informal published studies. Statistical power is the probability of detecting a meaningful difference between the variants when there really is one, that is, rejecting the null when there is a true difference of size δ. By setting the power to 80%, we can plug in a simple formula to determine the sample size—the number of users needed for each variant in an equally-sized design. The formula (van Belle 2002) is simply: $n = \frac{16\,\sigma^2}{\delta^2}$. The 16 in the numerator is designed for alpha of 0.05. Adjusting to 0.01 (see my spreadsheet) changes it as follows: $n = \frac{24\,\sigma^2}{\delta^2}$.

Let's assume that your conversion rate from visitor to purchaser is 5%, and you want to detect a 10% relative change. The σ^2 for a binomial metric is simply $p * (1 - p)$, so 0.05*0.95=0.0475. The 10% relative change is 0.05*10%=0.005, and that's the δ. Plug the numbers, and you have $n = \frac{24*0.0475}{0.005^2}$ = 45,600 users.

Where did the 10% come from? It is our pre-experiment estimate of the **minimum** relative change we want to detect. The smaller the number, the more users will be needed. I believe that 10% represents the upper bound for any reasonable non-triggered experiment (with triggering, you want the diluted effect to be at most 10%). The way to think about this is that if you **lost** this much, you might not get a statistically significant result. Ask your CFO if they would care if you launched a series of "flat" (not stat-sig) experiments that each potentially lost up to 10%; they would likely have a serious talk with you about how career-limiting that will be.

What if you had low power and got a stat-sig result? Maybe you're just betting on getting lucky? Well, no, it's called the winner's curse. With low power, your effect is likely to be highly exaggerated. Gelman and Carlin (2014) show that when power is below 50%, the exaggeration ratio, defined as the expectation of the absolute value of the estimate divided by the true effect size, becomes so high as to be meaningless, as shown in the figure below:



Exaggeration ratio as a function of statistical power (Gelman and Carlin 2014)

If you properly powered the above experiment and got a groundbreaking massive 20% improvement when you had 80% power to detect 10%, your p-value would be 0.00000000003, yes, that's 3E-11. Plugging in 100% improvement, as some of these examples claim, the p-value underflows to zero. The fact that with such massive improvements their p-values are around 0.01-0.05 implies that they are grossly underpowered.

Remaining Conditions

The last conditions specified above are easier to meet.

- Clear specification of the OEC (Overall Evaluation Criterion).
  Are we looking at conversion from the given page to the next, or to the last stage of the funnel, usually a purchase? It is sometimes easy to increase the micro-conversion of step X, which then hurts the conversion of the next step.
- All differences between the treatment(s) and control must be clearly stated.
  Ideally, a single factor is changed, that is OFAT, or One Factor At a Time. The problem with a treatment that changes three things is that it's impossible to know which one made the difference. In the first example above, the coupon code was removed, **and** a longer form was used. It's very possible that the removal of coupon code helped and the longer form hurt, and we only see the combination.
  If there is a performance difference (speed) or page weight (new page adds 5K to the page and makes it slower), these must be explicitly stated.
- Guardrails evaluated: in particular, at least a test for SRM (Sample-Ratio-Mismatch) must be done.
  The SRM test is like a seatbelt for your car: it detects a family of common scenarios where the result is invalid. There may be other tests that raise the trust level (e.g., fraud level shouldn't change when revenue increases).

Acknowledgment

Thanks to Deborah O'Malley for discussions about trust, and earlier discussions with Lukas Vermeer and Georgi Georgiev about statistical power.

References

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2017. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6-10. https://www.nature.com/articles/s41562-017-0189-z.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641 –651. doi:10.1177/1745691614551642.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, etal. 2019. "Top Challenges from the first Practical Online Controlled Experiments Summit." 21 (1). https://bit.ly/ControlledExperimentsSummit1.

Ioannidis, John P. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. doi:10.1371/journal.pmed.0020124.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251). doi:https://doi.org/10.1126/science.aac4716.

Simons, Daniel J. 2014. "The Value of Direct Replication." *Perspectives on Psychological Science* 91 (1): 76-80. doi:0.1177/1745691613514755.

van Belle, Gerald. 2002. *Statistical Rules of Thumb.* Wiley-Interscience.