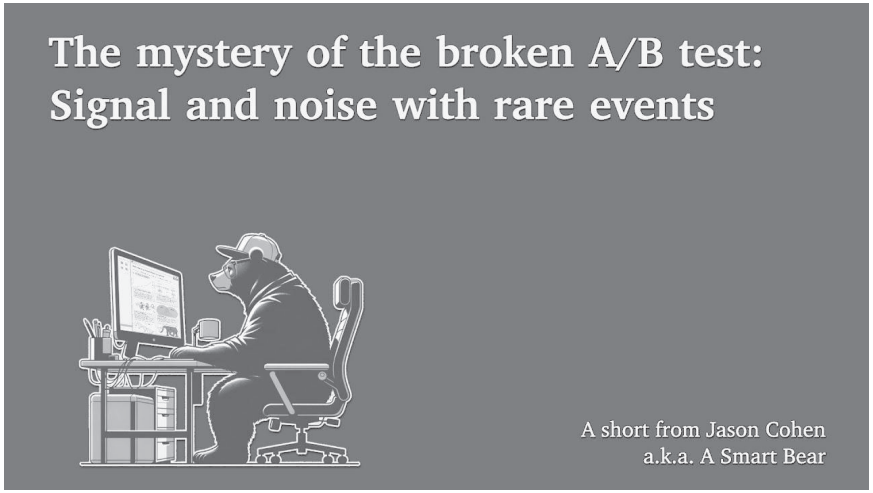
Easy statistics for A/B testing and hamsters

Determining whether an A/B test is statistically significant · Appendix for the mathematically inclined: The derivation



Watch on YouTube[1]

This video explains the concept, as well as the statistical principle that explains the mental fallacy that tricks us when we reason about rare events.

So you've got your AdWords test all set up: Will people go for the headline "Code Review Tools" or "Tools for Code Review?"

Gee they're both so exciting! Who could choose! I know, I know, settle down. Welcome to A/B testing.

Anyway, the next day you have this result:

	Variant A "Code Review Tools"	**Variant B "Tools for Code Review"**
Clicks:	31	19

Is this conclusive? Has A won? Or should you let the test run longer? Or should you try completely different text?

The answer matters. If you wait too long between tests, you're wasting time. If you don't wait long enough for *statistically conclusive* results, you might *think* a variant is better and use that false assumption to create a new variant, and so forth, all on a wild goose chase! That's not just a waste of time, it also prevents you from doing the *correct* thing, which is to come up with a *completely new test*.

Normally a formal statistical treatment would be too difficult, but I'm here to rescue you with a statistically sound yet incredibly simple formula that determines whether your A/B test results really are significant.

I'll get to it in a minute, but I can't help but include a more entertaining example than AdWords. Meet Hammy the Hamster, the probably-biased-but-incredibly-lovable tester of organic produce:

Watch Hammy the hamster on YouTube[2]

In the movie, Hammy chooses the organic produce **8 times** and the conventional **4 times**. This is an A/B test, just like with AdWords… but healthier.

If you're like me, you probably think "organic" is the clear-cut winner —after all Hammy chose it *twice as often* as conventional veggies. But, as so often happens with probability and statistics, **you'd be wrong**.

That's because human beings are notoriously bad at guessing these things from gut feel. Most people are more afraid of dying in a plane crash than a car crash, even though the latter is 1000x more likely.[3] On the other hand, we're amazed when CNN "calls the election" for a governor with a mere 1% of the state ballots reporting in. We also can't distinguish between patterns and noise.[4]

Okay okay, we suck at math. So what's the answer? Here's the bit you've been waiting for:*

* For the mathematical derivation, see the end of the article.

DETERMINING WHETHER AN A/B TEST IS STATISTICALLY SIGNIFICANT

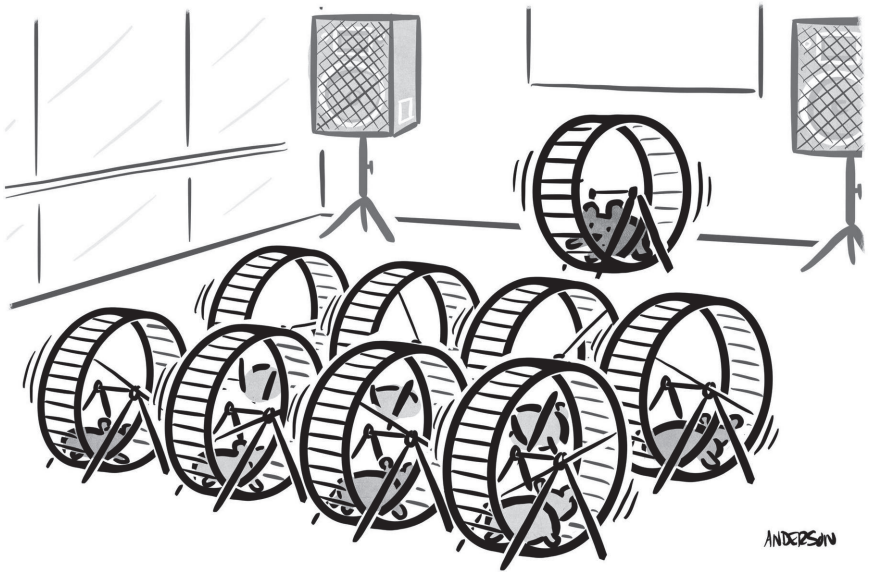
1. Define N as "the number of trials."
   For Hammy, $N = 8 + 4 = \mathbf{12}$
   For AdWords, $N = 31 + 19 = \mathbf{50}$
2. Define D as "half the difference between the 'winner' and the 'loser.'"
   For Hammy, $D = \dfrac{8 - 4}{2} = \mathbf{2}$

   For AdWords, $D = \dfrac{31 - 19}{2} = \mathbf{6}$
3. The test result is statistically significant only if $D^2 > N$.
   For Hammy, $D^2 = 4$, which is *not* bigger than 12, so it is *not significant*.
   For AdWords, $D^2 = 36$, which is *not* bigger than 50, so it is *not significant*.

So your AdWords test isn't statistically significant yet. But you let the test continue to run. The next day you find 31 more clicks for variant A, and 19 more clicks for B. Rerunning the test, the measured difference is now significant:

	Variant A	**Variant B**	N	D	D^2	**Stat Sig?**
Day one:	31	19	50	6	36	No
Day two:	62	38	100	12	144	Yes

A lot of times, though, you keep running the test and it's still not significant. That's when you realize you're not learning anything new; the



credit[5]

variants you picked are not meaningfully different for your readers. That means it's time to come up with something new.

When you start applying the formula to real-world examples, you'll notice that **when N is small, it is difficult—or even impossible—to be statistically significant**. For example, say you've got one ad with 6 clicks and the other with 1. That's $N = 7; D = 2.5; D^2 = 6.25$. So the test is still inconclusive, even though A is beating B six-to-one. Trust the math here—with only a few data points, you really don't know anything yet.

The smaller the N, the bigger the difference needs to be, to be detectable. Specifically, results are significant only if the ratio between A and B is larger than $\frac{\sqrt{N}+2}{\sqrt{N}-2}$. So for example, if $N = 50$, as it was in our "day one A/B test" example, the winning variant needs to be almost double the number of clicks as the losing variant in order to have a detectable difference, e.g. a conversation rate of 10% versus 5%. This is a huge difference; it's great if you find something so dramatically better, but this is rare, and therefore you can almost never find a significant A/B test given only 50 clicks to analyze.

When $N = 100$, the winner needs to be at least 50% higher than the loser (which it was by the second day in our example). It takes $N = 1800$ to detect the case where the winner is only 10% larger than the loser.

And this is bad news for A/B tests, because often one variant isn't better than the other by more than 10%, e.g. a "2.4% conversion rate" versus a "2.2% conversion rate." What does this mean, especially if you don't have large N? It means **you need to be seeking big differences**, not subtle ones. Test wildly different designs, rather than tweaks. Tweaks can only be tested when N is enormous.

I hope this formula will help you make the right choices when running A/B tests. It's simple enough that you have no excuse not to apply it! Human intuition sucks when it comes to these things, and A/B testing tools often use misleading or incorrect math, so let this formula help you draw the right conclusions.

APPENDIX FOR THE MATHEMATICALLY INCLINED: THE DERIVATION

The null-hypothesis[6] is that the results of the A/B test are due to chance alone. The statistical test we need is Pearson's chi-squared.[7]*

The definition of the χ^2 statistic follows, where:

m = number of possible outcomes;

O_k = observed quantity of results in category k;

E_k = expected quantity of results in category k:

* Not the Student t-test as is commonly claimed by people online who have only a passing familiarity with statistics; the t-test is appropriate with continuous, normally-distributed random variables, whereas χ^2 is appropriate for categorical random variables from independent trials and arbitrary probability distributions, which is what an A/B test is.

$$\chi^2 = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k}$$

In the simple case of a two-variant A/B test, $m = 2$. O_1 and O_2 are the observed results, and definitionally $N = O_1 + O_2$. The expected result under the null-hypothesis is that the quantities fall equally into each category, therefore $E_1 = E_2 = N/2$.

Plugging this into the definition:

$$\chi^2 = \frac{(O_1 - \frac{N}{2})^2}{\frac{N}{2}} + \frac{(O_2 - \frac{N}{2})^2}{\frac{N}{2}}$$

The first numerator can be rewritten in terms of O_1 and O_2 by substituting $N = O_1 + O_2$, and this results in our variable D^2 as defined in the main text:

$$\begin{aligned}
\left(O_1 - \frac{N}{2}\right)^2 &= \left(O_1 - \frac{O_1 + O_2}{2}\right)^2 \\
&= \left(\frac{2O_1 - O_1 - O_2}{2}\right)^2 \\
&= \left(\frac{O_1 - O_2}{2}\right)^2 \\
&= D^2
\end{aligned}$$

We can repeat with the second numerator, and so the expression simplifies:

$$\begin{aligned}
\chi^2 &= \frac{(O_1 - \frac{N}{2})^2}{\frac{N}{2}} + \frac{(O_2 - \frac{N}{2})^2}{\frac{N}{2}} \\
&= \frac{D^2}{\frac{N}{2}} + \frac{D^2}{\frac{N}{2}} \\
&= \frac{2}{N}(2D^2) \\
&= \frac{4D^2}{N}
\end{aligned}$$

Now that we have a simple formula for the chi-squared statistic, we refer to the chi-squared distribution to determine statistical significance. Specifically: What is the probability this result would have happened by chance alone?

Looking at the distribution[8] at 1 degree of freedom, we must exceed 3.8 for 95% confidence and 6.6 for 99% confidence. For this simplified rule-of-thumb formula, I selected 4 as the critical threshold. Solving for D^2 completes the derivation:

$$\chi^2 > 4$$

$$\frac{4D^2}{N} > 4$$

$$D^2 > N$$

□

(And if D^2 is more than double N, you're well past the 99% confidence level.)

Deriving the other statement in the article—that the ratio between the two variants needs to exceed a certain threshold to be significant—start with the boundary condition of being significant, and derive the values of A and B in that case:

$$
\begin{aligned}
D^2 &= N \\
D &= \sqrt{N} \\
A - \frac{N}{2} &= \sqrt{N} \\
A &= \frac{N}{2} + \sqrt{N}
\end{aligned}
$$

Similarly, given that $B = N - A$:

$$B = \frac{N}{2} - \sqrt{N}$$

And so:

$$
\begin{aligned}
\frac{A}{B} &= \frac{\frac{N}{2} + \sqrt{N}}{\frac{N}{2} - \sqrt{N}} \\
&= \frac{N + 2\sqrt{N}}{N - 2\sqrt{N}} \\
&= \frac{\sqrt{N} + 2}{\sqrt{N} - 2}
\end{aligned}
$$

The current version of this article:

https://asmartbear.com/ab-testing-statistics/

More articles & socials:

https://asmartbear.com

References

1. https://youtu.be/FaUO2-AQmr0
2. https://youtu.be/8z8CWdRaQpw
3. https://injuryfacts.nsc.org/all-injuries/preventable-death-overview/odds-of-dying/
4. https://longform.asmartbear.com/pattern-seeking-fallacy/
5. https://andertoons.com/exercise/cartoon/8888/hamster-exercise-wheel-class
6. https://en.wikipedia.org/wiki/Null_hypothesis
7. https://en.wikipedia.org/wiki/Pearson's_chi-square_test
8. https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm