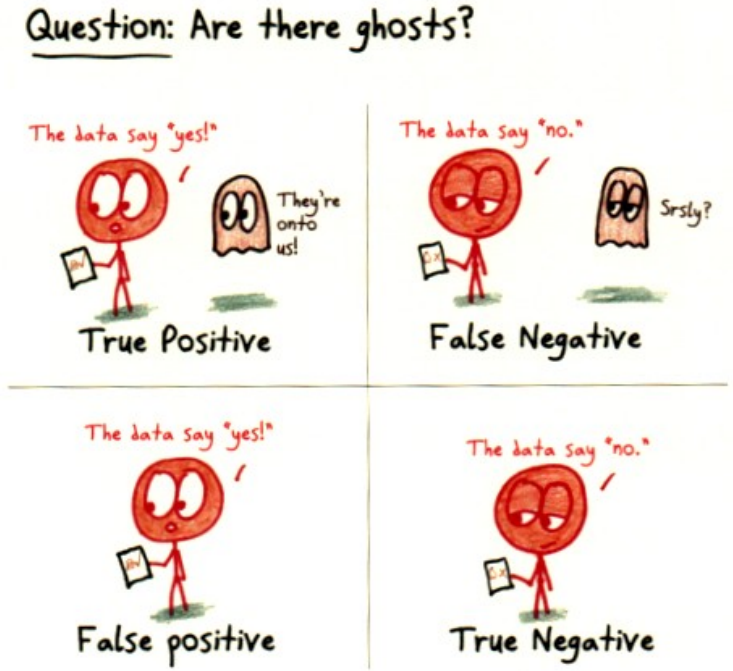
BARBARIANS AT THE GATE OF SCIENCE

THE P-VALUE CRISIS

Across the sciences, we are living through an epoch of crisis. After decades of the best science they could muster, many scholars now find their life's work hanging in the balance. The culprit isn't dishonesty, or a lack of integrity, or too many passages arguing against free will. The illness runs deeper, all the way down to a single statistic at the heart of the research process. It's a figure that made modern science possible – and that now threatens its stability.

1. TO CATCH A FLUKE

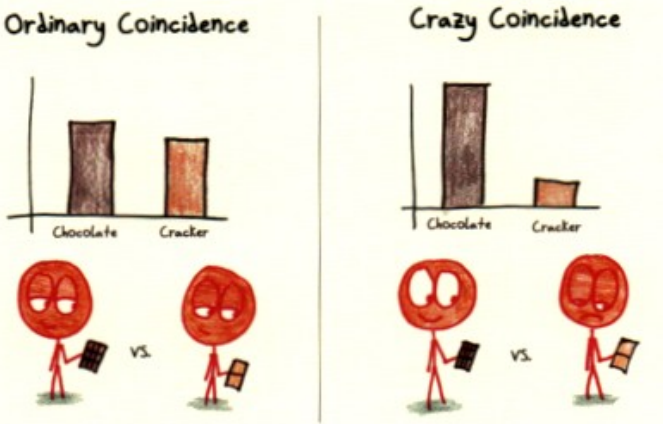
Every science experiment asks a question. Are gravitational waves real? Do millennials hate solvency? Can this new drug cure antivax paranoia? No matter the question, there are two possible truths ("yes" and "no") and, given the inherent unreliability of evidence, two possible outcomes ("you get it right" and "you get it wrong"). Thus, experimental results can fall into four categories:



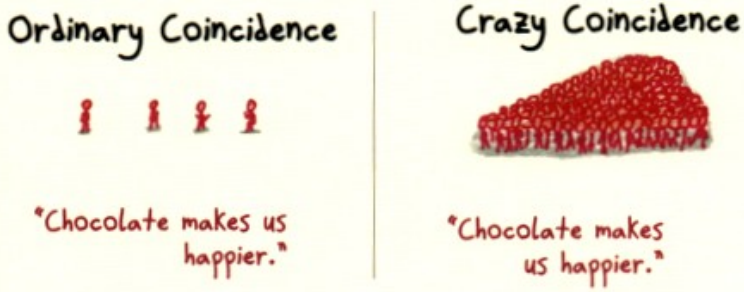
Scientists want **true positives**. They are known as "discoveries" and can win you things like Nobel Prizes, smooches from your romantic partner, and continued funding.

To weed out flukes, the p-value incorporates three fundamental factors:

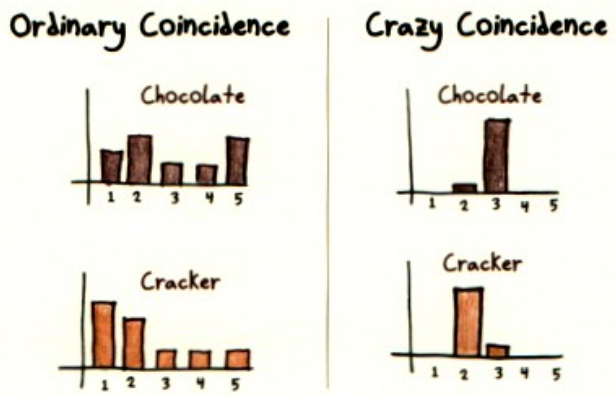
1. **The size of the difference.** A razor-thin margin (say, 3.3 vs. 3.2) is much likelier to occur by coincidence than a substantial gap is (say, 4.9 vs. 1.2).



2. **How big the data set is.** A two-person sample inspires little confidence. Maybe I happened to give the chocolate to an enthusiastic lover of life, and the graham cracker to an ungrateful nihilist. But in a randomly divided sample of two *thousand* people, individual differences should wash out. Even a smallish gap (3.08 vs. 3.01) is unlikely to happen by fluke.

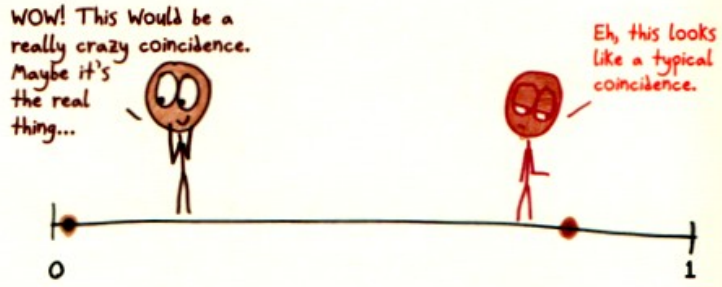


3. **The variance within each group.** When the scores are wild and high-variance, it's easy for two groups to get different results by fluke. But if the scores are consistent and low-variance, then even a small difference is hard to achieve by coincidence.



The p-value boils all of this information down into a single number between zero and 1, a sort of "coincidence craziness score." The lower the value, the crazier it would be for these results to happen by coincidence alone. A p-value near zero signals a coincidence so crazy that perhaps it isn't a coincidence at all.

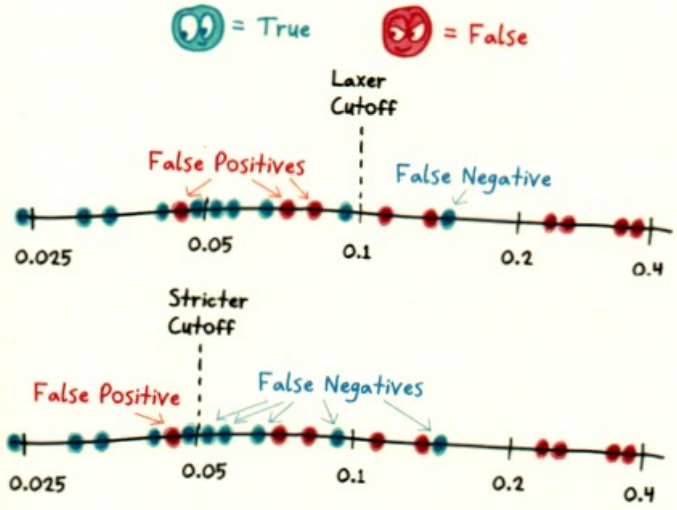
(For a slightly more technical discussion, see the endnotes.)



2. CALIBRATING THE FLUKE FILTER

In 1925, a statistician named R. A. Fisher published a book called *Statistical Methods for Research Workers*. In it, he proposed a line in the sand: 0.05. In other words, let's filter out 19 of every 20 flukes.

Why let through the other one in 20? Well, you can set the threshold lower than 5% if you like. Fisher himself was happy to consider 2% or 1%. But this drive to avoid false positives incurs a new risk: false *negatives*. The more flukes you weed out, the more true results get caught in the filter as well.

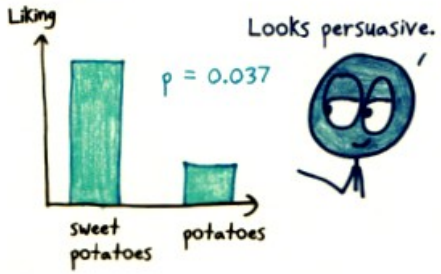


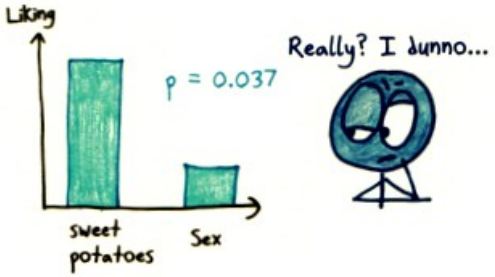
Suppose you're studying whether men are taller than women. Hint: they are. But what if your sample is a little fluky? What if you happen to pick taller-than-typical women and shorter-than-typical men, yielding an average difference of just 1 or 2 inches? Then a strict p-value threshold may reject the result as a fluke, even though it's quite genuine.

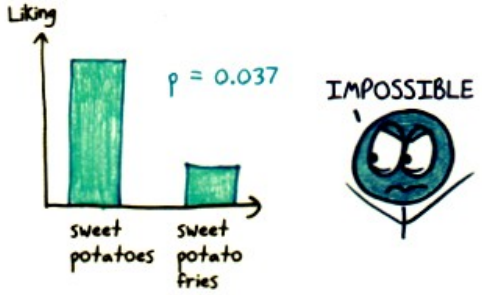
The number 0.05 represents a compromise, a middle ground between incarcerating the innocent and letting the guilty walk free.

For his part, Fisher never meant 0.05 as an ironclad rule.

All new evidence must be weighed against existing knowledge. Not all 0.04s are created equal.





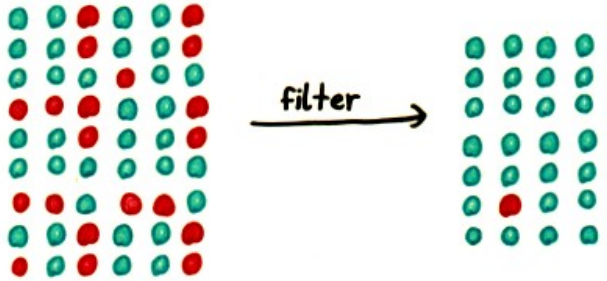


Scientists get this. But in a field that prides itself on standardization and objectivity, nuanced case-by-case judgments are hard to defend. And so, as the 20th century wore on, in human sciences like psychology and medicine the 5% line evolved from "suggestion" to "guideline" to "industry standard." p = 0.0499? Significant. p = 0.0501? Sorry, better luck next time.

Does this mean that 5% of certified results are flukes? Not quite. The reality is the reverse: 5% of flukes can become certified results. If that sounds equivalent, it's not.

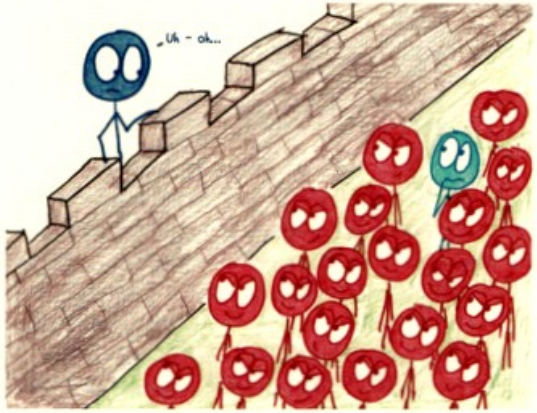
It's much scarier.

Imagine the p-value as a guardian in the castle of science. It wants to welcome the true positives inside the walls while repelling the barbarian false positives at the gate. We know that 5% of barbarians will slip through, but on balance, this seems good enough.

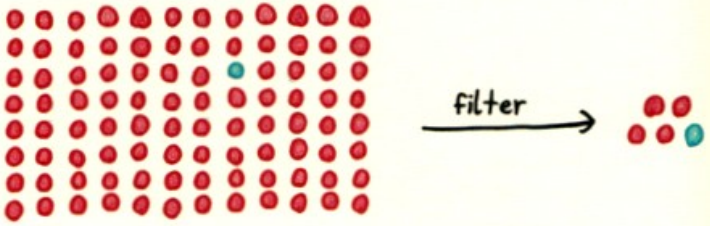


However, what if the attacking barbarians outnumber our own troops 20 to one? Five percent of the invading forces will equal the entirety of the civilized ones.

Worse yet, what if there are a *hundred* barbarians for every honest soldier? Their 5% will overwhelm the entire civilized army. The castle will teem with false positives, while the true ones cower in corners.



The danger, then, lies in scientists running too many studies where the real answer is "no." Does lip-syncing make your hair blonder? Can wearing clown shoes cause acid rain? Run a million junk studies, and 5% will clear the filter. That's 50,000. They'll flood the scientific journals, splash across headlines, and make Twitter even less readable than usual.



If that's not discouraging enough, it gets worse. Without meaning to, scientists have equipped the barbarians with grappling hooks and battering rams.

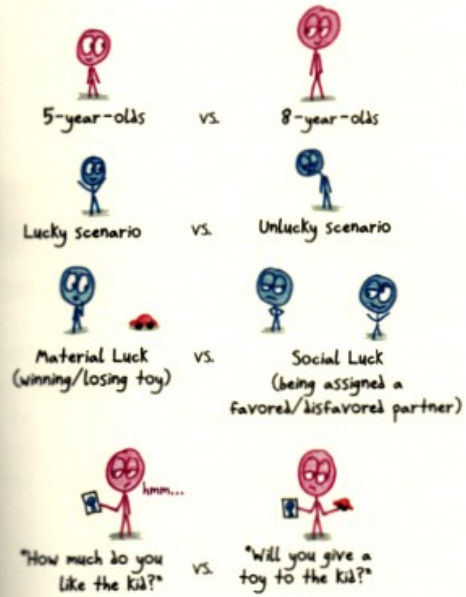
3. HOW FLUKES BREED

In 2006, a psychologist named Kristina Olson began to document a peculiar bias: Children prefer lucky people to unlucky ones. Olson and her collaborators found the bias across cultures, from age three to adulthood. It applied to those suffering little mishaps (like falling in the mud) or big calamities (like hurricanes). The effect was robust and persistent—a true positive.

Then, in 2008, Olson agreed to advise the senior thesis of a feckless 21-year-old named "me." With her copious help, I devised a modest follow-up, exploring whether five- and eight-year-olds would give more toys to the lucky than the unlucky.

I tested 46 kids. The answer was "no."

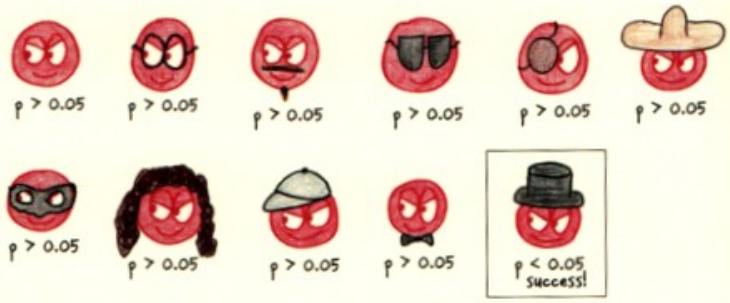
If anything, the reverse held: my subjects seemed to give more to the unlucky than the lucky. Far from "fun science fact time," this felt obvious; of course you give a toy to someone who lost one. Needing to squeeze 30 pages out of the experience, I looked to my data. Each subject had answered eight questions, and I had tested a variety of conditions. Thus, I could carve up the numbers in several ways:



By all appearances, my thesis was a barbarian at the gates. The crucial p-value was well above 0.05. But with an open mind, other possibilities emerged. What if I considered just the five-year-olds? Or just the eight-year-olds? Or just the lucky recipients? Or just the unlucky ones? Did gender make a difference? What if eight-year-old girls proved more context-sensitive than five-year-old boys in giving to children whom they rated at least 4 on the six-point "liking" scale?

What if, what if, what if . . .

By slicing and reslicing my data, I could transform one experiment into 20. It didn't matter if the p-value rejected my barbarian once, twice, or 10 times. I could equip it with new disguises until it finally snuck through into the castle.

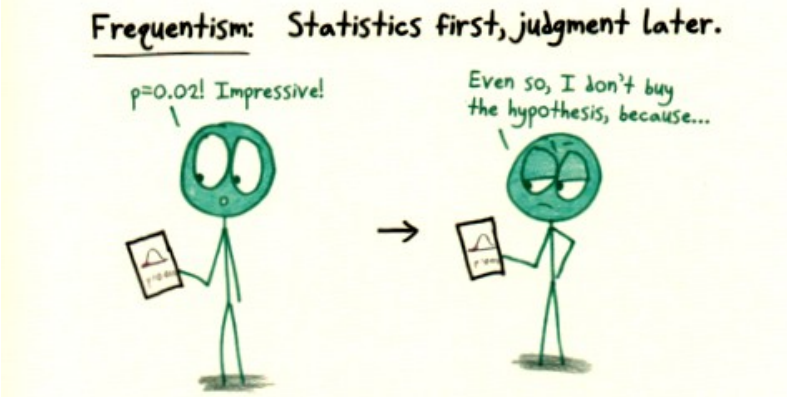


Thus is born perhaps the greatest methodological crisis of the young century: p-hacking. Take a bunch of truth-loving scientists, place them in a winner-take-all competition to achieve positive results, and then watch as, in spite of themselves, they act like 21-year-old me, rationalizing dodgy decisions. "Well, maybe I could run the numbers again . . . " "I *know* this result is real; I just need to exclude those outliers . . . " "Ooh, if I control for the 7th variable, the p-value drops to 0.03 . . . " Most research is ambiguous, with a mess of variables and a multitude of defensible ways to interpret the data. Which are you going to pick: the method that brands your results "insignificant" or the one that nudges them below 0.05?

4. THE WAR ON FLUKES

The replication crisis has rekindled an old rivalry between two gangs of statisticians: the frequentists and the Bayesians.

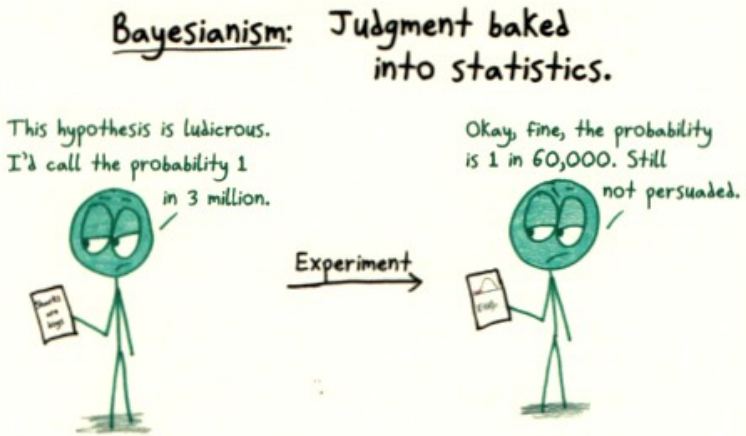
Ever since Fisher, frequentists have reigned. Their statistical models aim for neutrality and minimalism. No judgment calls. No editorializing. The p-value, for instance, doesn't care whether it's testing a surefire hypothesis or a mad-scientist one. That kind of subjective analysis comes later.



Bayesians reject this impartiality. Why should statistics feign indifference between plausible hypotheses and absurd ones, as if all 0.05's are created equal?

The Bayesian alternative works something like this. You begin with a "prior": an estimate of your hypothesis's probability. "Mints cure bad breath"? High. "Mints cure bad bones"? Low. You bake this estimate into the mathematics itself, via a rule called Bayes's formula. Then, after you run the experiment, statistics help you update the prior, weighing the new evidence against your old knowledge.

Bayesians don't care whether the results clear some arbitrary fluke filter. They care whether the data suffice to persuade us, whether they move the needle on our prior beliefs.



The Bayesians feel that their time has come. The frequentist regime, they assert, has led to ruin, and it's time to inaugurate a new era. Frequentists reply that priors are too arbitrary, too vulnerable to abuse. They propose their own reforms, such as lowering the p-value threshold from 0.05 (or 1 in 20) to 0.005 (or 1 in 200).