How Sure Is Sure Enough?

Why a finding that makes headlines in medicine wouldn't count as a discovery in physics

In July 2012, physicists at CERN announced the Higgs boson only after the chance of a statistical fluke had fallen to roughly one in 3.5 million. A medical journal will often publish a finding when that chance is one in twenty. Both get called "significant." What differs is the bar behind the word, by a factor of more than a hundred thousand, and the gap tells you a great deal about how much to trust the next study you read about.

What "significant" really means

Imagine a trial of a new blood-pressure pill. The treated group does a little better than the untreated group. Is that real, or the kind of difference luck throws up even when the pill does nothing? A p-value measures one narrow thing: how often you'd see a gap this big if the pill did nothing at all. A small p-value means the result would be surprising in a world of pure chance, which gives you reason to doubt that world. A large one means luck explains it easily.

A significance threshold is just a line drawn in advance. The familiar $p < 0.05$ came from Ronald Fisher in 1925, offered as a convenient marker rather than a law of nature. Clear the line and the result earns the label.

The trap is what the label doesn't mean. A p-value is *not* the probability your finding is true, and not the chance the pill does nothing. It's calculated by assuming the pill does nothing, so it can't also report how likely that assumption is. The American Statistical Association judged this misreading common enough to warn against it directly in 2016. Flip the logic anyway and you've made the most popular mistake in science reporting.

Two questions, not one

The mistake has a precise shape. Significance testing and Bayesian analysis answer different questions, and they're easy to confuse because each one mentions the data and a hypothesis.

> **Significance testing asks:**
>
> $$P(\text{data this extreme} \mid \text{the null is true})$$
>
> **Bayesian analysis asks:**
>
> $$P(\text{the hypothesis} \mid \text{data})$$
>
> **Bayes' theorem links the two:**
>
> $$P(\text{hypothesis} \mid \text{data}) = P(\text{data} \mid \text{hypothesis}) \times P(\text{hypothesis}) \div P(\text{data})$$
>
> *The jump from the first question to the second needs the prior, P(hypothesis).*

Read a small p-value as the chance your result is real and you've quietly swapped the first question for the second. Statisticians call it the transposed conditional, the same slip a court makes when it mistakes the odds of the evidence given innocence for the odds of innocence

given the evidence. Crossing between them takes a prior, the very piece a bare p-value never supplies. That swap is the subject of Aubrey Clayton's 2021 book *Bernoulli's Fallacy*.

Why physics asks for so much more

Physicists didn't choose five standard deviations, "five sigma," to be dramatic. A brand-new particle is unlikely to begin with, a false discovery can embarrass a field for a decade, and an experiment that scans countless possibilities will turn up flukes unless the bar sits high. A demanding threshold quietly does the work of the skeptical prior nobody writes down. Berger and Sellke showed in 1987 that a p-value just under 0.05 carries far less evidential weight than its face value suggests.

Context	Threshold	Odds of a fluke this big
A typical biomedical study	$p < 0.05$	about 1 in 20
Proposed stricter default (2017)	$p < 0.005$	about 1 in 200
Physics discovery (five sigma)	$p \approx 3 \times 10^{-7}$	about 1 in 3.5 million
Genome-wide gene studies	$p < 5 \times 10^{-8}$	about 1 in 20 million

The last column is the chance of a fluke this large if nothing real is going on, *not* the chance the finding is false.

Medicine isn't uniformly lax, either. Genome-wide studies, which test millions of genetic variants at once, insist on $p < 5 \times 10^{-8}$, stricter than the physics discovery bar, for the same reason physics is strict: test enough things and chance alone will hand you winners. And no drug reaches the market on a single significant trial. The FDA has long asked for more than one.

What to do with a "significant" result

Treat a lone significant result as a hint, not a verdict, and grow more skeptical the more surprising the claim. Run twenty studies of a treatment that does nothing and, on average, one will clear $p < 0.05$ by chance. Publish only that one and the record shows a victory built on noise. Multiply that across the millions of studies run worldwide, add the habit of publishing hits and burying misses, and you arrive at the picture John Ioannidis painted in his bluntly titled 2005 paper, "Why Most Published Research Findings Are False." In 2017 a large group of statisticians proposed simply moving the default bar to 0.005. Until something like that takes hold, the skepticism is yours to bring.

Sources

ATLAS Collaboration (2012). *Physics Letters B*, 716, 1–29; CMS Collaboration (2012). *Physics Letters B*, 716, 30–61.

Benjamin, D. J., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis. *Journal of the American Statistical Association*, 82, 112–122.

Clayton, A. (2021). *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. Columbia University Press.

Dudbridge, F., & Gusnanto, A. (2008). *Genetic Epidemiology*, 32, 227–234; Pe'er, I., et al. (2008). *Genetic Epidemiology*, 32, 381–385.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

U.S. Food and Drug Administration (1998). *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values. *The American Statistician*, 70, 129–133.