Lydia Denworth is a contributing editor for *Scientific American* and is author of *Friendship: The Evolution, Biology, and Extraordinary Power of Life's Fundamental Bond* (W. W. Norton, in press).

STATISTICS

A Significant Problem

Standard scientific methods are under fire. Will anything change?

By Lydia Denworth

In 1925 British geneticist and statistician Ronald Fisher published a book called *Statistical Methods for Research Workers*. The title doesn't scream "best seller," but the book was a huge success and established Fisher as the father of modern statistics. In it, he tackles the problem of how researchers can apply statistical tests to numerical data to draw conclusions about what they have found and determine whether it is worth pursuing. He references a statistical test that summarizes the compatibility of data with a proposed model and produces a *p* value. Fisher suggests that researchers might consider a *p* value of 0.05 as a handy guide: "It is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not." Pursue results with *p* values below that threshold, he advises, and do not spend time on results that fall above it. Thus was born the idea that a value of *p* less

IN BRIEF

The use of *p* values for nearly a century to determine statistical significance of experimental results has contributed to an illusion of certainty and reproducibility crises in many scientific fields.

There is growing determination to reform statistical analysis, but researchers disagree on whether it should be tweaked or overhauled. Some suggest changing statistical methods, whereas others would do away with a threshold for defining "significant" results.

Ultimately the *p* value plays into the human need for certainty. So it may be time for both scientists and the public to embrace the discomfort of being unsure.

Illustration by Mark Allen Miller

than 0.05 equates to what is known as statistical significance—a mathematical definition of "significant" results.

Nearly a century later, in many fields of scientific inquiry, a p value less than 0.05 is considered the gold standard for determining the merit of an experiment. It opens the doors to the essentials of academia—funding and publication—and therefore underpins most published scientific conclusions. Yet even Fisher understood that the concept of statistical significance and the p value that underpins it has considerable limitations. Most have been recognized for decades. "The excessive reliance on significance testing," wrote psychologist Paul Meehl in 1978, "[is] a poor way of doing science." P values are regularly misinterpreted, and statistical significance is not the same thing as practical significance. Moreover, the methodological decisions required in any study make it possible for an experimenter, consciously or unconsciously, to shift a p value up or down. "As is often said, you can prove anything with statistics," says statistician and epidemiologist Sander Greenland, professor emeritus at the University of California, Los Angeles, and one of the leading voices for reform. Studies that rely only on achieving statistical significance or pointing out its absence regularly result in inaccurate claims—they show things to be true that are false and things to be false that are true. After Fisher had retired to Australia, he was asked whether there was anything in his long career he regretted. He is said to have snapped, "Ever mentioning 0.05."

In the past decade the debate over statistical significance has flared up with unusual intensity. One publication called the flimsy foundation of statistical analysis "science's dirtiest secret." Another cited "numerous deep flaws" in significance testing. Experimental economics, biomedical research and especially psychology have been engulfed in a controversial replication crisis, in which it has been revealed that a substantial percentage of published findings are not reproducible. One of the more notorious examples is the idea of the power pose, the claim that assertive body language changes not just your attitude but your hormones, which was based on one paper that has since been repudiated by one of its authors. A paper on the economics of climate change (by a skeptic) "ended up having almost as many error corrections as data points—no kidding!—but none of these error corrections were enough for him to change his conclusion," wrote statistician Andrew Gelman of Columbia University on his blog, where he regularly takes researchers to task for shoddy work and an unwillingness to admit the problems in their studies. "Hey, it's fine to do purely theoretical work, but then no need to distract us with data," Gelman wrote.

The concept of statistical significance, though not the only factor, has emerged as an obvious part of the problem. In the past three years hundreds of researchers have urgently called for reform, authoring or endorsing papers in prestigious journals on redefining statistical significance or abandoning it altogether. The American Statistical Association (ASA), which put out a strong and unusual statement on the issue in 2016, argues for "moving to a world beyond $p < 0.05$." Ronald Wasserstein, the ASA's executive director, puts it this way: "Statistical significance is supposed to be like a right swipe on Tinder. It indicates just a certain level of interest. But unfortunately, that's not what statistical significance has become. People say, 'I've got 0.05, I'm good.' The science stops."

The question is whether anything will change. "Nothing is new. That needs to sober us about the prospect that maybe this time will be the same as every other time," says behavioral economist

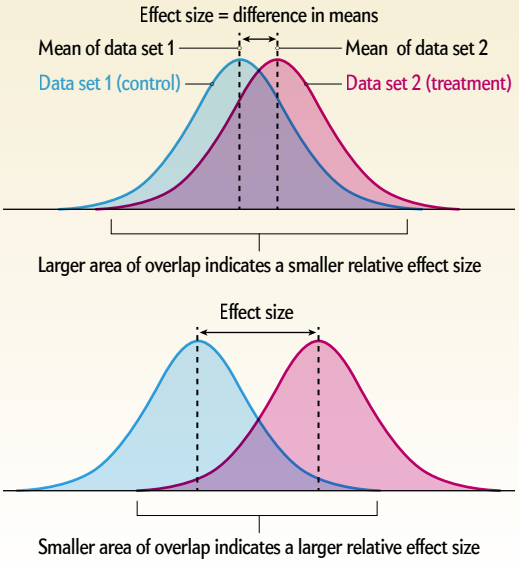
Statistical Significance

Imagine you grow pumpkins in your garden. Would using fertilizer affect their size? Given your long experience without fertilizer, you know how much the weights of pumpkins vary and you know that their average weight is 10 pounds. You decide to grow a sample of 25 pumpkins with fertilizer. The average weight of these 25 pumpkins turns out to be 13.2 pounds. How do you decide whether the difference of 3.2 pounds from the status quo of 10 pounds—the hypothetical "null" value—happened by chance or that fertilizer does indeed grow larger pumpkins?

Statistician Ronald Fisher's solution to this puzzle involves performing a thought experiment: imagine that you were to repeatedly grow 25 pumpkins a very large number of times. Each time you would get a different average weight because of the random variability of individual pumpkins. Then you would plot the distribution of those averages and consider the probability (**p value**) that the data you have generated would be possible if the fertilizer had no effect. By convention, a p value of 0.05 became a cut-off to identify significant results—in this case, ones that lead a researcher to conclude the fertilizer does not have an effect. Here we break down some of the concepts that drive the thought experiment for statistical significance.

EFFECT SIZE

The effect size for a treatment is the difference between the average outcome when the treatment is used compared with the average when the treatment is not used. The concept can be used to compare averages in samples or "true" averages for entire distributions. The effect size can be measured in the same units (such as pounds of pumpkins) as the outcome. But for many outcomes—such as responses to some psychological questionnaires—there is not a natural unit. In that case, researchers can use relative effect sizes. One way of measuring relative effect size is based on the overlap between the control and the treatment distributions.

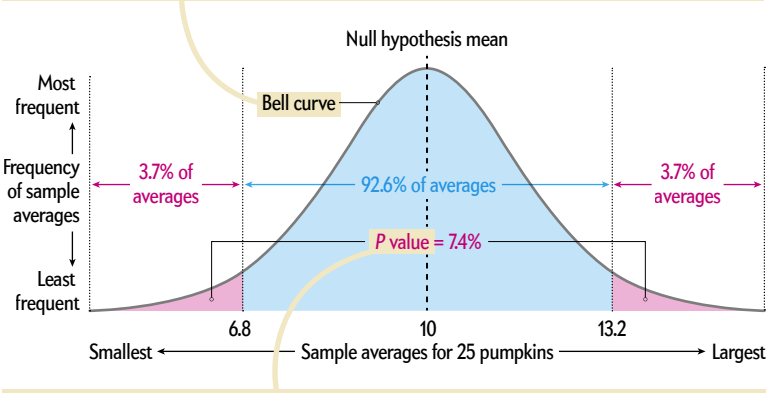


Graphic by Amanda Montañez (graphs) and Heather Krause

P VALUE

To calculate the *p* value, we need to compare the actual average of 13.2 pounds that we observed in our sample of 25 pumpkins with the random distribution of averages if we were to take many new samples of 25 pumpkins.

The bell curve shows the distribution of random average weights for samples of 25 under the null hypothesis that the fertilizer has no effect.

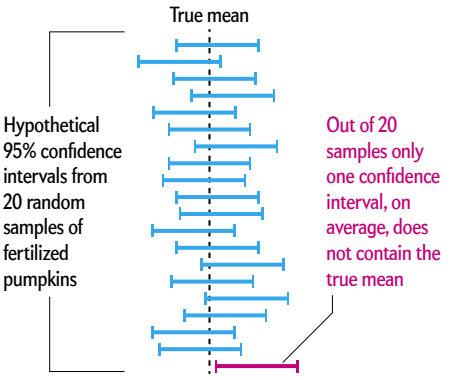


The *p* value is the probability of getting a random average weight as far from 10 as the average you actually observed, 13.2. Since 13.2 – 10 = 3.2, we want the probability of getting an average ≥ 13.2 or ≤ 6.8 (6.8 = 10 – 3.2). In this example, that probability is 0.074, which is the actual observed *p* value for your sample. Because it is greater than 0.05, your result would not be considered significant evidence that the fertilizer makes a difference.

The example shows a "two-tailed test," where the *p* value counts the probability of a weight greater than 13.2 and that of a weight less than 6.8 (10 – 3.2 = 6.8). Under some circumstances, a researcher might choose to perform a "one-tailed test." In that case, the *p* value would be only 0.037, which, being less than 0.05, is considered significant. This illustrates one way in which researchers can modify their stated intention for a study to achieve different *p* values with exactly the same data.

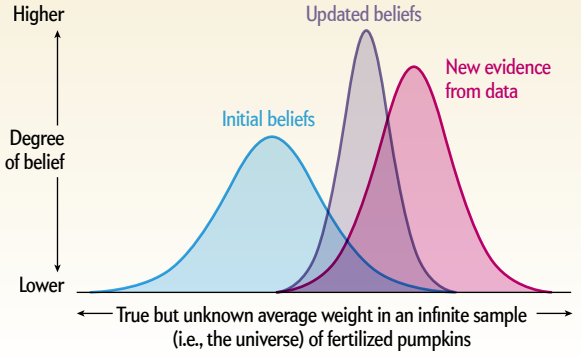
CONFIDENCE INTERVAL

We can calculate a 95 percent confidence interval from our sample of 25 pumpkins. This is a guess for the average weight of fertilized pumpkins. Calculating the 95 percent confidence interval involves inverting the calculation for the *p* value to find all hypothetical values that produce a *p* value ≥ 0.05. With our sample of 25 pumpkins, our 95 percent confidence interval goes from 9.69 to 16.71. The "true" average weight of fertilized pumpkins may or may not be in that interval. We can't be sure, so what does the "95 percent" mean? Imagine what would happen if we repeatedly grew batches of 25 pumpkins and sampled them. Each sample would produce a randomly different confidence interval. We know that in the long run, 95 percent of these intervals would include the true value and 5 percent would not. But what about our particular interval from the first pumpkin sample? We don't know whether it is in the 95 percent that worked or in the 5 percent that missed. It is the *process* that is right 95 percent of the time.



BAYESIAN METHODS

In the Bayesian approach to inference, a person's state of uncertainty about an unknown quantity is represented by a probability distribution. Bayes' theorem is used to combine individuals' initial beliefs—their distribution before looking at data—with the information they receive from the data, which produces a mathematically implied distribution for their updated beliefs. The updated beliefs from one study become the new initial beliefs for the next study, and so on. A major area of discussion and controversy concerns attempts to find "objective" criteria for initial beliefs. The goal is to find ways of constructing initial beliefs, known as prior distributions, that can be widely accepted by researchers as reasonable.



SURPRISAL

The *p* value conveys how surprising our pumpkin data are if we suppose that, in reality, fertilizing has no effect on growth. Some researchers have suggested that the *p* values do not convey surprisingness in a way that is intuitive for most people. Instead they suggest a mathematical quantity called a surprisal, also known as an *s* value or Shannon transform, that adjusts *p* values to produce bits (as in computer bits). Surprisal can be interpreted through the example of tossing coins.

Two heads in a row = **2** bits of surprisal = *p* value of $1/2^2$ = **0.25**

Four heads in a row = **4** bits of surprisal = *p* value of $1/2^4$ = **0.0625**

Five heads in a row = **5** bits of surprisal = *p* value of $1/2^5$ = **0.03215**

Our sample of 25 pumpkins with an average weight of 13.2 and a *p* value of 0.074 produces between 3 and 4 bits of surprisal. To be exact: 3.76 bits of surprisal since $3.76 = -\log_2(0.074)$.

Daniel Benjamin of the University of Southern California, another voice for reform. Still, although they disagree over the remedies, it is striking how many researchers do agree, as economist Stephen Ziliak wrote, that "the current culture of statistical significance testing, interpretation, and reporting has to go."

THE WORLD AS IT IS

THE GOAL OF SCIENCE is to describe what is true in nature. Scientists use statistical models to infer that truth—to determine, for instance, whether one treatment is more effective than another or whether one group differs from another. Every statistical model relies on a set of assumptions about how data are collected and analyzed and how the researchers choose to present their results.

Those results nearly always center on a statistical approach called null hypothesis significance testing, which produces a *p* value. This testing does not address the truth head-on; it glances at it obliquely. That is because significance testing is intended to indicate only whether a line of research is worth pursuing further. "What we want to know when we run an experiment is how likely is it [our] hypothesis is true," Benjamin says. "But [significance testing] answers a convoluted alternative question, which is, if my hypothesis were false, how unlikely would my data be?"

Sometimes this works. The search for the Higgs boson, a particle first theorized by physicists in the 1960s, is an extreme but useful example. The null hypothesis was that the Higgs boson did not exist; the alternative hypothesis was that it must exist. Teams of physicists at CERN's Large Hadron Collider ran multiple experiments and got the equivalent of a *p* value so vanishingly small that it meant the possibility of their results occurring if the Higgs boson did not exist was one in 3.5 million. That made the null hypothesis untenable. Then they double-checked to be sure the result wasn't caused by an error. "The only way you could be assured of the scientific importance of this result, and the Nobel Prize, was to have reported that [they] went through hoops of fire to make sure [none] of the potential problems could have produced such a tiny value," Greenland says. "Such a tiny value is saying that the Standard Model without the Higgs boson [can't be correct]. It's screaming at that level."

But physics allows for a level of precision that isn't achievable elsewhere. When you're testing people, as in psychology, you will never achieve odds of one in three million. A *p* value of 0.05 puts the odds of repeated rejection of a correct hypothesis across many tests at one in 20. (It does not indicate, as is often believed, that the chance of error on any single test is 5 percent.) That's why statisticians long ago added "confidence intervals," as a way of providing a sense of the amount of error or uncertainty in estimates made by scientists. Confidence intervals are mathematically related to *p* values. *P* values run from 0 to 1. If you subtract 0.05 from 1, you get 0.95, or 95 percent, the conventional confidence interval. But a confidence interval is simply a useful way of summarizing the results of hypothesis tests for many effect sizes. "There's nothing about them that should inspire any confidence," Greenland says. Yet over time both *p* values and confidence intervals took hold, offering the illusion of certainty.

P values themselves are not necessarily the problem. They are a useful tool when considered in context. That's what journal editors and scientific funders and regulators claim they do. The concern is that the importance of statistical significance might be exaggerated or overemphasized, something that's especially easy to do with

small samples. That's what led to the current replication crisis. In 2015 Brian Nosek, co-founder of the Center for Open Science, spearheaded an effort to replicate 100 prominent social psychology papers, which found that only 36.1 percent could be replicated unambiguously. In 2018 the Social Sciences Replication Project reported on direct replications of 21 experimental studies in the social sciences published in *Nature* and *Science* between 2010 and 2015. They found a significant effect in the same direction as in the original study for 13 (62 percent) of the studies, and the effect size of the replications was on average about half the original effect size.

Genetics also had a replication crisis in the early to mid-2000s. After much debate, the threshold for statistical significance in that field was shifted dramatically. "When you find a new discovery of a genetic variance related to some disease or other phenotype, the standard for statistical significance is 5×10^{-8}, which is basically 0.05 divided by a million," says Benjamin, who has also worked in genetics. "The current generation of human genetics studies is considered very solid."

The same cannot be said for biomedical research, where the risk tends toward false negatives, with researchers reporting no statistical significance when effects exist. The absence of evidence is not evidence of absence, just as the absence of a wedding ring on someone's hand is not proof that the person isn't married, only proof that the person isn't wearing a ring. Such cases sometimes end up in court when corporate liability and consumer safety are at stake.

BLURRING BRIGHT LINES

JUST HOW MUCH TROUBLE is science in? There is fairly wide agreement among scientists in many disciplines that misinterpretation and overemphasis of *p* values and statistical significance are real problems, although some are milder in their diagnosis of its severity than others. "I take the long view," says social psychologist Blair T. Johnson of the University of Connecticut. "Science does this regularly. The pendulum will swing between extremes, and you've got to live with that." The benefit of this round, he says, is

that it is a reminder to be modest about inferences. "If we don't have humility as scholars, we're not going to move forward."

To truly move forward, though, scientists must agree on solutions. That is nearly as hard as the practice of statistics itself. "The fear is that taking away this long-established practice of being able to declare things as statistically significant or not would introduce some kind of anarchy to the process," Wasserstein says. Still, suggestions abound. They include changes in statistical methods, in the language used to describe those methods and in the way statistical analyses are used. The most prominent ideas have been put forth in a series of papers that began with the ASA statement in 2016, in which more than two dozen statisticians agreed on several principles for reform. That was followed by a special issue of one of the association's journals that included 45 papers on ways to move beyond statistical significance.

In 2018 a group of 72 scientists published a commentary called "Redefine Statistical Significance" in *Nature Human Behaviour* endorsing a shift in the threshold of statistical significance from 0.05 to 0.005 for claims of new discoveries. (Results between 0.05 and 0.005 would be called "suggestive.") Benjamin, the lead author of that paper, sees this as an imperfect short-term solution but as one that could be implemented immediately. "My worry is that if we don't do something right away, we'll lose the momentum to do the kind of bigger changes that will really improve things, and we'll end up spending all this time arguing over the ideal solution. In the meantime, there will be a lot more damage that gets done." In other words, don't let the perfect be the enemy of the good.

Others say redefining statistical significance does no good at all because the real problem is the very existence of a threshold. In March, U.C.L.A.'s Greenland, Valentin Amrhein, a zoologist at the University of Basel, and Blakeley McShane, a statistician and expert in marketing at Northwestern University, published a comment in *Nature* that argued for abandoning the concept of statistical significance. They suggest that p values be used as a continuous variable among other pieces of evidence and that confidence intervals be renamed "compatibility intervals" to reflect what they actually signal: compatibility with the data, not confidence in the result. They solicited endorsements for their ideas on Twitter. Eight hundred scientists, including Benjamin, signed on.

Clearly, better—or at least more straightforward—statistical methods are available. Gelman, who frequently criticizes the statistical approaches of others, does not use null hypothesis significance testing in his work at all. He prefers Bayesian methodology, a more direct statistical approach in which one takes initial beliefs, adds in new evidence and updates the beliefs. Greenland is promoting the use of a surprisal, a mathematical quantity that adjusts p values to produce bits (as in computer bits) of information. A p value of 0.05 is only 4.3 bits of information against the null. "That's the equivalent to seeing four heads in a row if someone tosses a coin," Greenland says. "Is that much evidence against the idea that the coin tossing was fair? No. You'll see it occur all the time. That's why 0.05 is such a weak standard." If researchers had to put a surprisal next to every p value, he argues, they would be held to a higher standard. An emphasis on effect sizes, which speak to the magnitude of differences found, would also help.

Improved education about statistics for both scientists and the public could start with making the language of statistics more accessible. Back when Fisher embraced the concept of "significance," the word carried less weight. "It meant 'signifying' but not 'important,'" Greenland says. And it's not surprising that the term "confidence intervals" tends to instill undue, well, confidence.

EMBRACE UNCERTAINTY

STATISTICAL SIGNIFICANCE has fed the human need for certainty. "The original sin is people wanting certainty when it's not appropriate," Gelman says. The time may have come for us to sit with the discomfort of not being sure. If we can do that, the scientific literature will look different. A report about an important finding "should be a paragraph, not a sentence," Wasserstein says. And it shouldn't be based on a single study. Ultimately a successful theory is one that stands up repeatedly to decades of scrutiny.

Small changes are occurring among the powers that be in science. "We agree that p values are sometimes overused or misinterpreted," says Jennifer Zeis, spokesperson for the *New England Journal of Medicine*. "Concluding that a treatment is effective for an outcome if $p < 0.05$ and ineffective if $p > 0.05$ is a reductionist view of medicine and does not always reflect reality." She says their research reports now include fewer p values, and more results are reported with confidence intervals without p values. The journal is also embracing the principles of open science, such as publishing more detailed research protocols and requiring authors to follow prespecified analysis plans and to report when they deviate from them.

At the U.S. Food and Drug Administration, there hasn't been any change to requirements in clinical trials, according to John Scott, director of the Division of Biostatistics. "I think it's very unlikely that p values will disappear from drug development anytime soon, but I do foresee increasing application of alternative approaches," he says. For instance, there has been greater interest among applicants in using Bayesian inference. "The current debate reflects generally increased awareness of some of the limitations of statistical inference as traditionally practiced."

Johnson, who is the incoming editor at *Psychological Bulletin,* has seen eye to eye with the current editor but says, "I intend to force conformity to fairly stringent standards of reporting. This way I'm sure that everyone knows what happened and why, and they can more easily judge whether methods are valid or have flaws." He also emphasizes the importance of well-executed meta-analyses and systematic reviews as ways of reducing dependence on the results of single studies.

Most critically, a p value "shouldn't be a gatekeeper," McShane says. "Let's take a more holistic and nuanced and evaluative view." That was something that even Ronald Fisher's contemporaries supported. In 1928 two other giants of statistics, Jerzy Neyman and Egon Pearson, wrote of statistical analysis: "The tests themselves give no final verdict but as tools help the worker who is using them to form his final decision." SA

MORE TO EXPLORE

Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015. Colin F. Camerer et al. in *Nature Human Behaviour,* Vol. 2, pages 637–644; September 2018.

Moving to a World beyond "p< 0.05." Ronald L. Wasserstein, Allen L. Schirm and Nicole A. Lazar in *American Statistician,* Vol. 73, Supplement 1, pages 1–19; 2019.

FROM OUR ARCHIVES

Make Research Reproducible. Shannon Palus; October 2018.

scientificamerican.com/magazine/sa