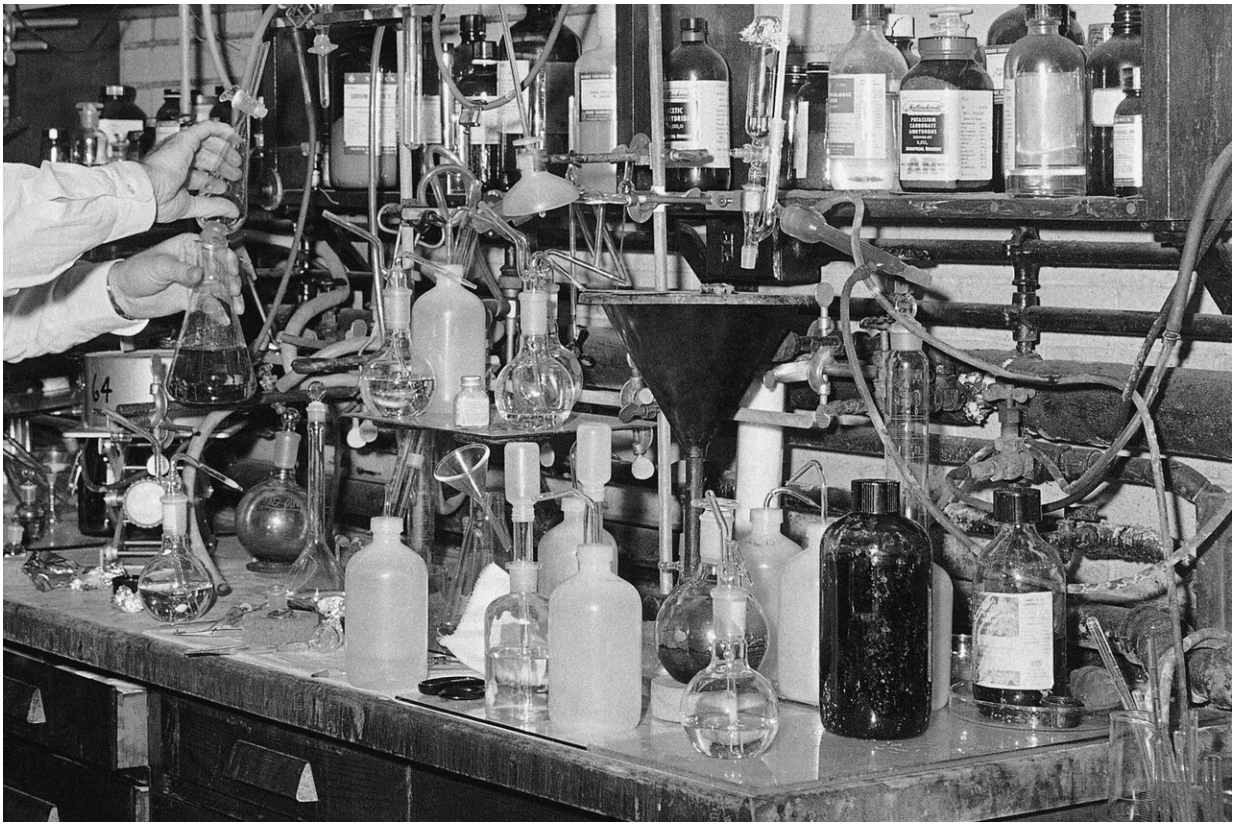
Sorry, wrong number: Statistical benchmark comes under fire

November 12 2019, by Malcolm Ritter



In this July 1, 1960 file photo, a chemist works in laboratory in Cambridge, Mass. For decades, scientists have used "statistical significance" to estimate whether their results are reliable or just flukes. It's long been criticized, but 2019 has brought two high-profile calls to get rid of it entirely. (AP Photo/Peter J. Carroll)

Earlier this fall Dr. Scott Solomon presented the results of a huge heart drug study to an audience of fellow cardiologists in Paris.

The results Solomon was describing looked promising: Patients who took the medication had a lower rate of hospitalization and death than patients on a different drug.

Then he showed his audience another number.

"There were some gasps, or 'Ooohs,'" Solomon, of Harvard's Brigham and Women's Hospital, recalled recently. "A lot of people were disappointed."

One investment analyst reacted by reducing his forecast for peak sales of the drug—by $1 billion.

What happened?

The number that caused the gasps was 0.059. The audience was looking for something under 0.05.

What it meant was that Solomon's promising results had run afoul of a statistical concept you may never have heard of: statistical significance. It's an all-or-nothing thing. Your statistical results are either significant, meaning they are reliable, or not significant, indicating an unacceptably high chance that they were just a fluke.

The concept has been used for decades. It holds a lot of sway over how scientific results are appraised, which studies get published, and what medicines make it to drugstores.

But this year has brought two high-profile calls from critics, including from inside the arcane world of statistics, to get rid of it—in part out of

concern that it prematurely dismisses results like Solomon's.

Significance is reflected in a calculation that produces something called a p-value. Usually, if this produces a p-value of less than 0.05, the study findings are considered significant. If not, the study has failed the test.

Solomon's study just missed. So the apparent edge his drug was showing over the other medication was deemed insignificant. By this criterion there was no "real" difference.

Solomon believes the drug in fact produced a real benefit and that a larger or longer-lasting study could have reached statistical significance.

"I'm not crying over spilled milk," he said. "We do set the rules. The question is, is that the right way to go about it?"

He's not alone in asking that question.

"It is a safe bet that people have suffered or died because scientists (and editors, regulators, journalists and others) have used significance tests to interpret results," epidemiologist Kenneth Rothman of RTI Health Solutions in Research Triangle Park, N.C., and Boston University wrote in 2016.

The danger is both that a potentially beneficial medical finding can be ignored because a study doesn't reach statistical significance, and a harmful or fruitless medical practice could be accepted simply because it does, he said in an email.

The p-value cutoff for significance Is "a measure that has gained gatekeeper status ... not only for publication but for people to take your results seriously," says Northwestern University statistician Blake McShane.

It's no wonder that a statistician, at a recent talk to journalists about the issue just before Halloween, displayed a slide of a jack-o'-lantern carved with this sight, obviously terrifying to anyone in science or medicine: "P = .06."

McShane and others argue that the importance of the p-value threshold is undeserved. He co-authored a call to abolish the notion of statistical significance, which was published in the prestigious journal Nature this year. The proposal attracted more than 800 co-signers.

Even the American Statistical Association, which had never issued any formal statement on specific statistical practices, came down hard in 2016 on using any kind of p-value cutoff in this way. And this year it went further, declaring in a special issue with 43 papers on the subject, "It is time to stop using the term "statistically significant' entirely."

What's the problem? McShane and others list several:

— P-value does not directly measure the likelihood that the outcome of an experiment just is a fluke. What it really represents is widely misunderstood, even by scientists and some statisticians, said Nicole Lazar, a statistics professor at the University of Georgia.

— Using a label of statistical significance "gives more certainty that is actually warranted," Lazar said. "We should recognize the fact that there is uncertainty in our findings."

— The traditional cutoff of 0.05 is arbitrary.

— Statistical significance does not necessarily mean "significant"—or that a finding is important practically or scientifically, Lazar says. It might not even be true: Solomon cites a large heart drug study that found a significant treatment effect for patients born in August but not July,

obviously just a random fluctuation.

— The term "statistical significance" sets up a goal line for researchers, a clear measure of success or failure. That means researchers can try a little bit too hard to reach it. They may deliberately game the system to get an acceptable p-value, or just unconsciously choose analytic methods that help, McShane and Lazar said.

— That can distort the effects not only of individual experiments, but also the cumulative results of studies on a given topic, so that overall a drug can look "a lot better than it actually is," McShane said.

What should be done instead? Abolish the bright line of statistical significance, and just report the p-value along with other analyses to give a more comprehensive outline of what the test result may mean, McShane and others say.

It may not be as clear-cut as a simple declaration of significance or insignificance, but "we'll have a better idea of what's going on," Lazar said. "I think it will be easier to weed out the bad work."

Not everybody buys the idea of doing away with statistical significance. Prominent Stanford researcher Dr. John Ioannidis says that abolition "could promote bias. Irrefutable nonsense would rule." Although he agrees that a p-value standard of less than 0.05 is weak and easily abused, he believes scientists should use a more stringent p-value or other statistical measure instead, specified before the experiment is performed.

McShane said that although calls for abolishing statistical significance have been raised for years, there seems to be more momentum lately.

"Maybe," he said, "it's time to put the nail in the coffin on this one for

good."