NATURE | NEWS

Weak statistical standards implicated in scientific irreproducibility

One-quarter of studies that meet commonly used statistical cutoff may be false.

Erika Check Hayden

11 November 2013

The plague of non-reproducibility in science may be mostly due to scientists' use of weak statistical tests, as shown by an innovative method developed by statistician Valen Johnson, at Texas A&M University in College Station.

Johnson compared the strength of two types of tests: frequentist tests, which measure how unlikely a finding is to occur by chance, and Bayesian tests, which measure the likelihood that a particular hypothesis is correct given data collected in the study. The strength of the results given by these two types of tests had not been compared before, because they ask slightly different types of questions.

So Johnson developed a method that makes the results given by the tests — the *P* value in the frequentist paradigm, and the Bayes factor in the Bayesian paradigm — directly comparable. Unlike frequentist tests, which use objective calculations to reject a null hypothesis, Bayesian tests require the tester to define an alternative hypothesis to be tested — a subjective process. But Johnson developed a 'uniformly most powerful' Bayesian test that defines the alternative hypothesis in a standard way, so that it "maximizes the probability that the Bayes factor in favor of the alternate hypothesis exceeds a specified threshold," he writes in his paper. This threshold can be chosen so that Bayesian tests and frequentist tests will both reject the null hypothesis for the same test results.

Johnson then used these uniformly most powerful tests to compare *P* values to Bayes factors. When he did so, he found that a *P* value of 0.05 or less — commonly considered evidence in support of a hypothesis in fields such as social science, in which non-reproducibility has become a serious issue — corresponds to Bayes factors of between 3 and 5, which are considered weak evidence to support a finding.

False positives

Indeed, as many as 17–25% of such findings are probably false, Johnson calculates[1]. He advocates for scientists to use more stringent *P* values of 0.005 or less to support their findings, and thinks that the use of the 0.05 standard might account for most of the problem of non-reproducibility in science — even more than other issues, such as biases and scientific misconduct.

"Very few studies that fail to replicate are based on *P* values of 0.005 or smaller," Johnson says.

Some other mathematicians said that though there have been many calls for researchers to use more stringent tests[2], the new paper makes an important contribution by laying bare exactly how lax the 0.05 standard is.

"It shows once more that standards of evidence that are in common use throughout the empirical sciences are dangerously lenient," says mathematical psychologist Eric-Jan Wagenmakers of the University of Amsterdam. "Previous arguments centered on '*P*-hacking', that is, abusing standard statistical procedures to obtain the desired results. The Johnson paper shows that there is something wrong with the *P* value itself."

Other researchers, though, said it would be difficult to change the mindset of scientists who have become wedded to the 0.05 cutoff. One implication of the work, for instance, is that studies will have to include more subjects to reach these more stringent cutoffs, which will require more time and money.

"The family of Bayesian methods has been well developed over many decades now, but somehow we are stuck to using frequentist approaches," says physician John Ioannidis of Stanford University in California, who studies the causes of non-reproducibility. "I hope this paper has better luck in changing the world."

Nature | doi:10.1038/nature.2013.14131

References

1. Johnson, V. E. *Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1313476110 (2013).

2. Ioannidis, J.P., Tarone, R. & McLaughlin, J. *Epidemiology*. **22**, 450–456 (2011).