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.

Top picks from nature news

- Humans interbred with a mysterious archaic population
- How the capacity to evolve can itself evolve
- The weak statistics that are making science irreproducible

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.

Related stories

- US behavioural research studies skew positive
- The data detective
- Replication studies: Bad copy

"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

Follow Erika on Twitter @Erika_Check.

References

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

2. Ioannidis, J.P., Tarone, R. & McLaughlin, J. *Epidemiology.* **22**, 450–456 (2011).
Show context Article PubMed ISI

Related stories and links

From nature.com

- **US behavioural research studies skew positive**
  26 August 2013
- **The data detective**
  03 July 2012
- **Replication studies: Bad copy**
  16 May 2012

For the best commenting experience, please login or register as a user and agree to our Community Guidelines. You will be re-directed back to this page where you will see comments updating in real-time and have the ability to recommend comments to other users.

13 comments

Subscribe to comments

Amelia Still • 2013-11-20 05:27 PM

Johnson makes the claim that "Very few studies that fail to replicate are based on P values of 0.005 or smaller," but I don't see any references or anywhere it is tested in the paper. Does anyone know where this claim comes from?

David Huffman • 2013-11-15 11:39 AM

Interesting, but does not address the main problem with contemporary western science, which is the encroachment of advocacy and the defense of conclusions into scientific thought, at the expense of healthy skepticism. Whether I'm a Bayesian or a frequentist, if I am thinking that the results of even the most rigorously designed experiment point to a specific interpretation or conclusion, I am in trouble to begin with. Probably the most effective way to reform scientific thought and discussion would be to discourage the use of such words as "evidence," "support," and any other such teleological language, and encourage more skeptical language that helps us to think critically about even our own conclusions. No experiment ever "supports," or "provides evidence for" a conclusion, and this article is replete with such misleading language. Indeed, whether I fall out on the frequentist side or the Bayesian side of this debate, the most sound scientific conclusion I can make regarding experimental data made reluctantly, and the conclusion I go with should recognize that it is merely the best explanation I can think of to explain my

results because a better explanation did not show up (yet) to displace it

Abhilash Dwarakanath • 2013-11-18 01:28 PM

Discourage evidence? Support? And, pray, what kind of sceptical language would you suggest that we use? As soon as someone talks of "problems with contemporary western science", its a red-flag.

David Huffman • 2013-11-20 01:12 PM

Sorry for the delay, but thanks for your question. For skeptical language that does not suggest that experimental results point to a specific conclusion, I would suggest that, when referring to our analysis of experimental results, we use words and phraseologies like "corroborates" and "is consistent with," because this language avoids the misleading teleological implications associated with the contemporary usage of words like "evidence" and "support." Please allow, for illustration, a simple example of a nondirectional t-test conducted on an experiment in which one random subset from the target population (the treatment group) has been exposed to a treatment that we hypothesize has been causing some observed but otherwise unexplained variation in a response variable, and another random subset (the control group) has been exposed to a placebo that we expect to have no effect on the response variable. After seeing the results, if the mean response of the treatment group is displaced substantially away from the mean of the control group, then the results are consistent with the prediction around which the experiment was designed. But before I call mom and tell her I am about to become famous, it would be prudent to address the first threat to the validity of my incipient conclusions — that threat being that the results could have occurred by random chance alone, without the assistance of any systematically applied phenomenological influence. In oder to address this issue, I must determine the likelihood that two piles of numbers with means this far apart could have arisen from a purely random process, and run a t-test to obtain that likelihood; which we call the “p” value. Convention (which is a relict of the day when we were stuck with mechanical calculators and tables of critical values) established that a useful line in the sand (alpha) would be p=0.05, so that there would be no more than a 5% likelihood that the experiment could have produced two groups of this size with means this far apart or farther. In declaring the null hypothesis an unreasonable sole explanation for the results, I do not confirm the phenomenological hypothesis around which my experiment was designed. No, I can only conclude that random sampling is not a plausible sole cause of the results, and therefore a more plausible conclusion is that at least one phenomenological agent was applied systematically to move the group means this far away from each other. But note, when I declare

the null to be an implausible sole cause, the more plausible alternate to the null is not the phenomenological hypothesis around which my experiment was designed, but is simply a conclusion that one or more phenomena were operating to contribute to the results. The probability that the null could be responsible for the results is p, and the probability that the alternate is a better conclusion is 1–p. however, the probability that the treatment variable contributed to the results is completely unknowable. We often forget that the null and its alternate are statistical hypotheses, and that both are “ignorant,” for lack of a better term, of the experimental design. Neither “knows” anything about the phenomenological hypothesis around which the experiment was designed. A successful t-test does not confirm anything about the experimental design or the effects of the treatment variable, but simply allows us to conclude that something phenomenological was probably operating through the experimental design to contribute to the results, but does not, in any way, finger the treatment variable as even one of the contributors to the non-random results. Indeed, the only way I can implicate the treatment variable as a contributor to the results is by a thorough review of my experimental design to see if there is any other plausible non-random contributor. Failing to find another cause, I then conclude, by default or forfeiture, that the only cause I can associate with the results is the treatment variable I used in an attempt to manipulate the responses of the treatment group. Therefore, the p-value with which I reject the null only affects the probability of the alternate (which is statistical, not phenomenological) being true, but has no bearing whatsoever on the probability that my treatment variable influenced the responses of the subjects in my treatment group. Consequently, any language which implies that the rejection of the null favors or supports my conclusion that the treatment variable influences the response variable is misleading at best. And may I suggest that by saying that the results of my experiment “were consistent with prediction” is also more consistent with the above logic than it would be if I said that my experiment “confirms” the phenomenological hypothesis around which I carefully designed my experiment.

Kenneth Pimple • 2013-11-13 07:52 PM

The apparent link in paragraph 4 reading "non-reproducibility has become a serious issue" seems to be a link to nothing. Is this an error or a commentary?

Tom Roberts • 2013-11-13 03:19 PM

In high energy physics (aka particle physics), the standard for calling something a discovery is generally five sigma, or P ~ 3E-7 for a normally distributed value. For instance, this was the standard used by CERN

in claiming discovery of the Higgs boson (there was more to it than just statistics). Five sigma is ENORMOUSLY less likely to be due to chance than a mere 5%. Indeed, we don't even bother to look at anything less than three sigma (P ~ 0.003, smaller than the suggested 0.005). With a criterion of a mere 5% I well believe irreproducibility is a problem -- in HEP nobody would take that seriously at all.

Jane Public • 2013-11-13 01:30 AM

If it requires more subjects in order to make the science valid and reproducible, then that is what it takes, yes? That is to say, what does in mean to complain that doing responsible science is "too difficult"? Is the person who is complaining arguing in favor of abdicating responsible science? I honestly do not understand arguments of that kind.

Michael Soljak • 2013-11-12 12:47 PM

Can anyone provide me with a doi for the Johnson article which works? The one at the bottom of this article doesn't, and I can't find the article on the Proc Nat Acad Sci website using Advanced Search either.

Emily Banham • 2013-11-12 03:08 PM

Dear Michael, I have found that the link above works ok, but here it is copied again in case a different source helps you access it. http://www.pnas.org/content/early/2013/10/28/1313476110

Andy Lawrence • 2013-11-11 08:44 PM

I suppose I should read the Johnson article... but this piece seems to suggest that P=0.05 itself is somehow suspect, and therefore Bayesian techniques are better than frequentist ones. Wuh? Surely it is simply that poor scientists underestimate errors and/or apply inappropriate distributions, and/or ignore multiple trials. Simple methodological screw-ups. There is nothing wrong with P=0.05 if you do the work right.

Nick Long • 2013-11-13 02:57 AM

Not true, it shows that P=0.05 is indeed an inappropriate threshold for accepting hypotheses in science, it IS suspect and causing an enormous waste of resources. And yes Bayesian techniques, if

available, are always better than frequentist ones.

Michael Weissman • 2013-11-11 11:01 PM

That's actually not what the article says. It's not another empirical summary of incorrectly done studies. It argues that for rather typical parameter estimation problems, the likelihood ratio for the alternate to null hypotheses is not nearly 19/1 for a p=0.05 cutoff. In other words, even for properly done studies there will typically be more false positives than people usually think. This is quite aside from the issue of overall prior probabilities, which can tilt either way.

Oliver H. • 2013-11-11 09:01 PM

If you have a certain probability of error and a sufficiently high number of publications, you will have plenty of errors out there. So reducing the probability of a type one error is certainly a way to reduce the number of publications that are flukes.

See other News & Comment articles from *Nature*

36 EISSN 1476-4687