NATURE | NEWS

Smart software spots statistical errors in psychology papers

One in eight articles contain data-reporting mistakes that affect their conclusions.

Monya Baker

28 October 2015

Researchers have created an automated software package that extracts and checks statistics from psychology research papers — and it has confirmed that the literature contains plenty of errors.

The program, statcheck, took less than 2 hours to crunch through more than 30,000 papers published in 8 respected journals between 1985 and 2013, says Michèle Nuijten, who studies analytical methods at Tilburg University in the Netherlands and is an author of a report on the work published in *Behavior Research Methods*[1].

The software found that 16,700 of the papers included tests of statistical significance in a format that it could check. Of these, it discovered that 50% had reported at least one *P* value — which reflects the likelihood of getting observed results if a null hypothesis is true — that was inconsistent with related statistical parameters in the test. And 13% of the papers contained an error that changed their conclusion: non-significant results were reported as significant, or vice versa.

These rates are in line with previous studies that assessed inconsistencies in *P* values manually, but checking by hand is tedious, says Nuijten. “On average, a psychology article contains about 11 *P* values. That means that statcheck checked it in the blink of an eye, whereas by hand it would take you about 10 minutes,” she says.

Overall, the software package looked at more than 258,000 *P* values reported in the papers, of which 10% were inconsistent with the related data. Most of those errors did not influence any statistical conclusion, because the *P* value was not near the 0.05 that is used as a benchmark of statistical significance. But the majority of the mistakes that did matter were non-significant *P* values wrongly presented as significant. Papers are not getting noticeably worse at weeding out errors, however: the average prevalence of inconsistent *P* values was stable, or may have declined slightly, between 1985 and 2013, the report found.

Large-scale literature checks

Daniele Fanelli, who studies scientific bias and behaviour at the Meta-Research and Innovation Center at Stanford University in Palo Alto, California, says that this is an important finding. “It suggests that, contrary to common speculations, the rate of errors and therefore potentially of sloppiness or questionable research practices have not grown in recent decades.” He thinks that the tool will be important mostly to meta-researchers analysing the literature on a large scale.

There are other software packages that can check *P* values when the data are plugged into them, but Nuijten is not aware of any others that can automatically extract values from papers. She thinks that her team’s program (written in the popular open-source software package R) could be applied easily by researchers, reviewers and editors. “I think a lot of reporting errors could be avoided if researchers used statcheck to check their own papers,” she says. “If journals already use an automated check for plagiarism, it seems a small step to include a statistical check.”

However, statcheck is not the final word, she warns; it occasionally wrongly interprets results as inconsistent, and it will miss results that are not reported in the standard American Psychological Association format used in most psychology papers.

The program is no panacea, says Eric-Jan Wagenmakers, a psychology researcher and statistician at the University of Amsterdam. “I think it is more useful to abandon *P* values altogether. But if you have to report *P* values, it can't hurt to use the program,” he says.

Nature | doi:10.1038/nature.2015.18657

References

1. Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S. & Wicherts, J. M. *Behav. Res.* http://dx.doi.org/10.3758/s13428-015-0664-2 (2015).

1 comment

Subscribe to comments

Daniel Lakens • 2015-10-29 01:12 PM

For an analysis of the kind of errors people make, see http://daniellakens.blogspot.nl/2015/10/checking-your-stats-and-some-errors-we.html. Most of the errors are very small differences in p-values, mainly due to the use of < instead of =. Formally, these alter the conclusion, but practically, they are of little consequence. Other common errors are the use of one-tailed test without clearly mentioning this, the incorrect use of p = 0.000, and incorrect rounding and copy paste errors.