TOOLBOX

A SPELLCHECKER FOR STATISTICS

Researchers debate whether using software to automatically detect inconsistencies in papers improves the literature, or raises false alarms.

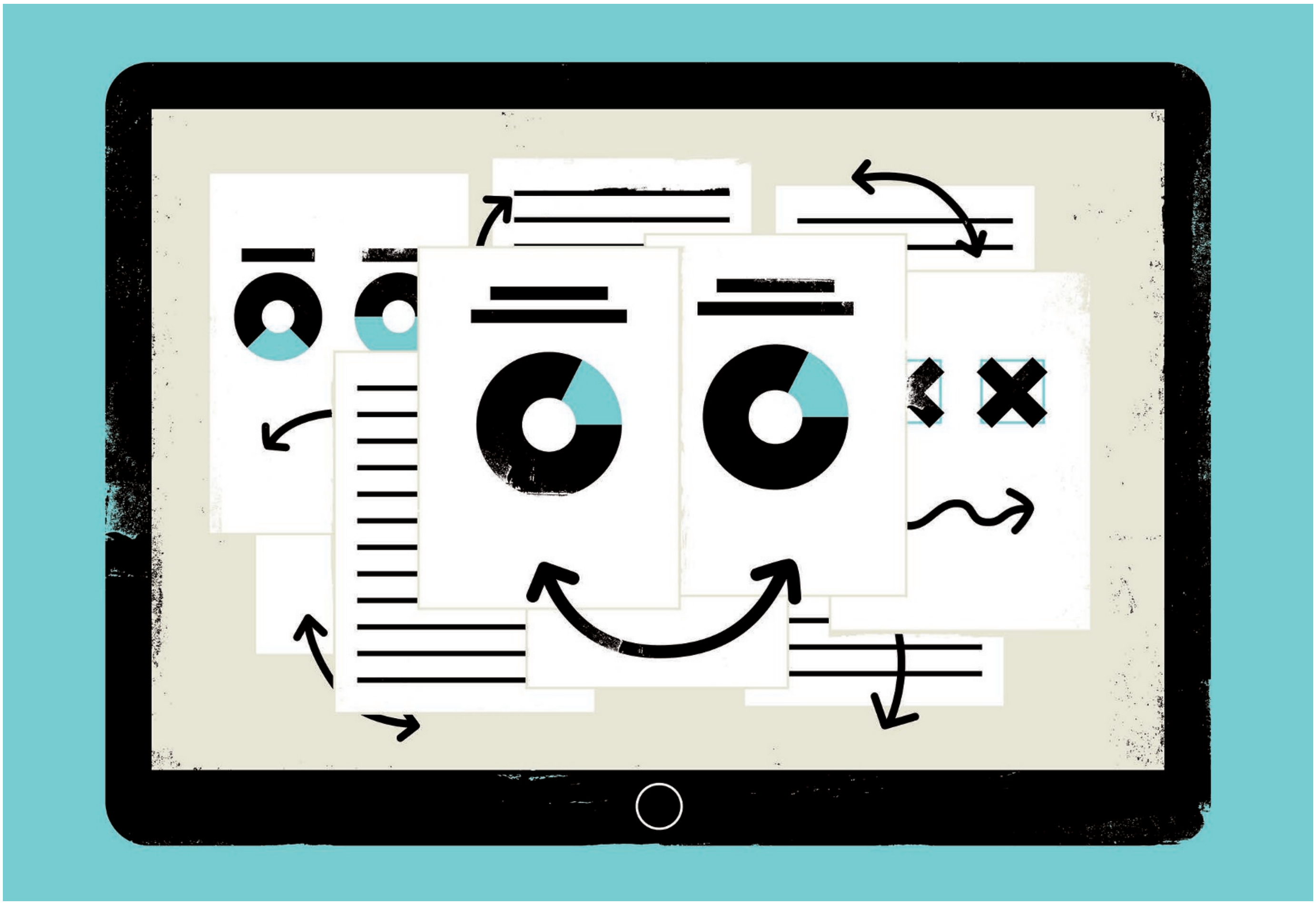


ILLUSTRATION BY THE PROJECT TWINS

BY MONYA BAKER

Michèle Nuijten and her colleagues found rampant inconsistencies when they unleased statcheck on the psychological literature. The program scans articles for statistical results, redoes the calculations and checks that the numbers match. It went through 30,717 papers to identify 16,695 that tested hypotheses using statistics. In half of those, it found at least one potential error (M. B. Nuijten *et al. Behav. Res. Methods* **48,** 1205–1226; 2016).

Nuijten did not alert the papers' authors. But this August, her co-author Chris Hartgerink, a fellow methodologist at Tilburg University in the Netherlands, moved the focus from the literature in general to specific papers. He set statcheck to work on more than 50,000 papers, and posted its reports on PubPeer, an online forum in which scientists often dispute papers. That has prompted a sometimes testy debate about how such tools should be used.

Hartgerink predicted that the posts would inform readers and authors about potential errors and "benefit the field more directly than just dumping a data set". Not everyone agreed. On 20 October, the German Psychological Association warned that posting false findings of error could damage researchers' reputations. And later that month, a former president of the Association for Psychological Science in Washington DC decried the rise of "uncurated, unfiltered denigration" through blogs and social media, and implied that posts from statcheck-like programs could be seen as harassment.

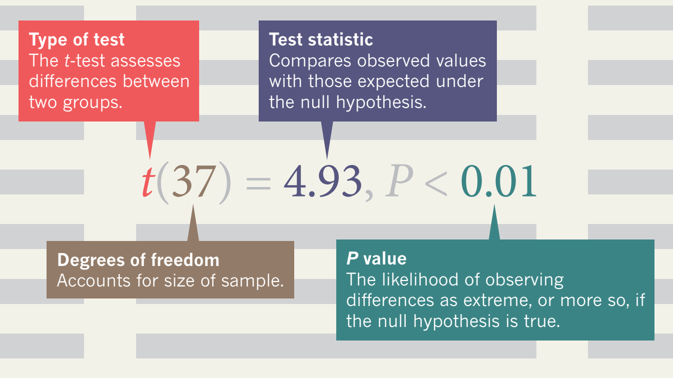
Others foresee a positive change in the culture. Hartgerink and Nuijten have each received awards from organizations promoting open science. And in a PubPeer comment ▶

▸ on the original statcheck paper, psychology researcher Nick Brown of the University of Groningen in the Netherlands wrote that science might benefit if researchers stopped assuming that posts on the forum indicated that there was “something naughty” in a paper, and instead thought, “There’s a note on PubPeer, I will read it and evaluate it like a scientist.”

An automated tool makes researchers more likely to double-check their work, which is good for psychology, argues Simine Vazire, who studies self-perception at the University of California, Davis. “It will catch mistakes, but even more importantly it will make us more careful.”

WHAT STATCHECK LOOKS FOR

This computer algorithm scans papers for statistical tests, uses reported results to recompute the *P* value and flags up inconsistencies.



$t(37) = 4.93, P < 0.01$

Type of test The *t*-test assesses differences between two groups.

Test statistic Compares observed values with those expected under the null hypothesis.

Degrees of freedom Accounts for size of sample.

***P* value** The likelihood of observing differences as extreme, or more so, if the null hypothesis is true.

That seems to appeal. Several thousand people have downloaded the free statcheck program, which works in the programming language R, or visited the web-based statcheck.io, which requires no programming knowledge. (Researchers who want to check selected results rather than whole papers can use online calculators such as ShinyApps.)

TECHNICAL CHECK

Most psychology papers report statistical tests in a standardized format, with related parameters that can be checked for inconsistencies. Statcheck — which so far works only for papers in this format — identifies and inspects a few common tests that calculate *P* values, a measure of how likely results are to arise by chance if, for instance, no real difference exists between two groups (see ‘What statcheck looks for’). Although statisticians have warned against it, a *P* value below 0.05 is often used as an arbitrary determiner of ‘statistical significance’, allowing results to be taken seriously and published.

Most of the errors that statcheck catches seem to be typos or copy-and-pasting mistakes, says Daniel Lakens, a cognitive psychologist at Eindhoven University of Technology in the Netherlands. After reading the statcheck paper, he decided to analyse the errors it reported that changed a result’s statistical significance. He found three main categories. Often, a researcher had inserted an incorrect sign, such as $P < 0.05$ instead of $P = 0.05$. In other cases, the calculations were set up to detect only particular relationships, such as positive or negative correlation, which was not always made explicit. Optimistic rounding was also common: *P* values of 0.055 reported as $P \leq 0.05$ made up 10% of detected errors that changed statistical significance, a rate that Lakens calls depressingly high.

But statcheck itself makes errors, says Thomas Schmidt, an experimental psychologist at the University of Kaiserslautern in Germany, who wrote a critique of the program (T. Schmidt Preprint at http://arxiv.org/abs/1610.01010; 2016) after it flagged two of his papers. For example, it does not always recognize necessary statistical adjustments.

When statcheck does detect an error, it cannot distinguish whether it is the *P* value or a related parameter that is incorrect. Schmidt says that, across the two of his papers that it scanned, statcheck failed to detect 43 *P* values, checked 137 and noted 35 “potentially incorrect statistical results”. Of those, 2 reflected *P*-value errors that did not change significance, 3 reflected errors in other parameters but did not affect *P* values, and 30 were improperly flagged.

Nuijten admits that statcheck can sometimes misidentify tests and overlook adjusted *P* values, but she notes that, in her original paper, it found similar rates of error to manual checks.

Nuijten and Hartgerink have been working hard, mostly successfully, to keep conversations amiable. Nuijten has posted detailed explanations about how statcheck works, with smiley emoji and friendly exclamation marks. Hartgerink is updating PubPeer posts with an improved version of the software. Both note that anyone can add comments on PubPeer to explain statcheck’s results, and that the posts state that results are not definitive. “The one thing I try to repeat over and over is that statcheck is automated software that will never be as accurate as a manual check,” says Nuijten.

Much of what statcheck flags up is trivial, but when authors do not respond, matters are left unresolved, says Elkan Akyürek, a psychologist at the University of Groningen. “Content-based discussion is getting a bit flooded.” Thought leaders such as neuropsychologist Dorothy Bishop of the University of Oxford, UK, worry that posts could distract from more serious discussions, or alienate people and make them less receptive to efforts to improve reproducibility. Heiko Hecht, a psychologist at Johannes Gutenberg University in Mainz, Germany, thinks it might have the opposite effect: “The program is still very immature, but in the long run could keep scientists honest.” Besides, he adds, if researchers made raw data available, anyone could check the results.

Some authors have expressed gratitude for a chance to correct mistakes, although several have said that they should have the chance to review posts before they are made public. At least three have responded on PubPeer to explain errors. Two of them told *Nature* that the errors were typos that did not affect *P* values and were too trivial to justify a formal correction. As for Vazire, she hopes that automated reports will help researchers to get used to post-publication commentary. “I think it will help desensitize us to criticism,” she says.

EDITOR’S HELPER

In July this year, the journal *Psychological Science* began running statcheck on submissions that got favourable first reviews, and discussing flagged inconsistencies with the authors. “I thought there might be some blowback or resistance,” says editor-in-chief Stephen Lindsay. “Reaction has been almost non-existent.” Of the few dozen runs so far, none of the errors has been egregious, he says, although there have been at least two instances in which authors have reported a *P* value as 0.05 when it was 0.054.

Lindsay says that statcheck reports are too confusing to share with authors directly. (For example, the program flags potential errors with the word TRUE.) Nuijten says that an upcoming version will be much more comprehensible to non-programmers. Meanwhile, she says, her team has been talking to publishers Elsevier and PLOS about adopting the program at their titles. And statcheck may soon have company: a more-comprehensive commercial program called StatReviewer is under development by other researchers. It is designed to analyse papers from a variety of fields, not just to double-check calculations but also to ensure that reporting requirements are followed.

Lindsay hopes that statcheck’s utility will fade over time as researchers stop manually entering statistical outcomes into their manuscripts; instead, the values would be directly inserted by the programs that produced them, and linked to their scripts. “The methodological leaders are using things like R markdown,” he says.

As for Schmidt, he thinks that statcheck could be useful in manuscript preparation, but it is not for beginners. “The greatest risk during prepublication is that unsophisticated users overestimate the program, relying blindly on its output.” Lakens is sticking to a manual system: one author of a paper does the analyses, and another checks them. That can detect errors that statcheck will not, such as transposing results.

That approach makes sense to Nuijten. Her goal was never to fix statistical analysis. Statcheck is more like a standard spellchecker, she says: “a handy tool that sometimes says stupid things”. People laugh at the absurdities, but still use the tool to correct mistakes. ■