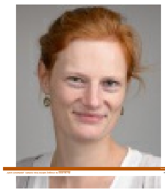
statcheck – a spellchecker for statistics



A study has revealed a high prevalence of inconsistencies in reported statistical test results. Such inconsistencies make results unreliable, as they become "irreproducible", and ultimately affect the level of trust in scientific reporting. statcheck is a free, open-source tool that automatically extracts reported statistical results from papers and recalculates p-values. Following an investigation into its accuracy, ***Michèle B. Nuijten*** *finds statcheck to be very effective at flagging inconsistencies and gross inconsistencies, with an overall accuracy of 96.2% to 99.9%.*

If you're a non-native English speaker (like me), but you often have to write in English (like me), you will probably agree that the spellchecker is an invaluable tool. And even when you do speak English fluently, I'm sure that you've used the spellchecker to filter out any typos or other mistakes.

When you're writing a scientific paper, there are many more things that can go wrong than just spelling. One thing that is particularly error-prone is the reporting of statistical findings.

Statistical errors in published papers

Unfortunately, we have plenty of reasons to assume that copying the results from a statistical program into a manuscript doesn't always go well. Published papers often contain impossible means, coefficients that don't add up, or ratios that don't match their confidence intervals.

In psychology, my field, we found a high prevalence of inconsistencies in reported statistical test results (although these problems are by no means unique to psychology). Most conclusions in psychology are based on "null hypothesis significance testing" (NHST) and look roughly like this:

> "The experimental group scored significantly higher than the control group, $t(58) = 1.91$, $p < .05$".

This is a t-test with 58 degrees of freedom, a test statistic of 1.91, and a *p*-value that is smaller than .05. A *p*-value smaller than .05 is usually considered "statistically significant".

This example is, in fact, inconsistent. If I recalculate the *p*-value based on the reported degrees of freedom and the test statistic, I would get $p = .06$, which is not statistically significant anymore. In psychology, we found that roughly half of papers contain at least one inconsistent *p*-value, and in one in eight papers this may have influenced the statistical conclusion.

Even though most inconsistencies we found were small and likely to be the result of innocent copy-paste mistakes, they can substantively distort conclusions. Errors in papers make results unreliable, because they become "irreproducible": if other researchers would perform the same analyses on the same data, a different conclusion would roll out. This, of course, affects the level of trust we place in these results.

statcheck

The inconsistencies I'm talking about are obvious. Obvious, in the sense you don't need raw data to see that certain reported numbers don't match. The fact that these inconsistencies do arise in the literature means that peer review did not filter them out. I think it could be useful to have an automated procedure to flag inconsistent numbers. Basically, we need a spellchecker for stats. To that end, we developed statcheck.

statcheck is a free, open-source tool that automatically extracts reported statistical results from papers and recalculates *p*-values. It is available as an R package and as a user-friendly web app at http://statcheck.io.

statcheck roughly works as follows. First, it converts articles to plain-text files. Next, it searches the text for statistical results. This is possible in psychology, because of the very strict reporting style (APA); stats are always reported in the same way. When statcheck detects a statistical result, it uses the reported degrees of freedom and test statistic to recompute the *p*-value. Finally, it compares the reported *p*-value with the recalculated one, to see if they match. If not, the result is flagged as an inconsistency. If the reported *p*-value is significant and the recalculated one is not, or vice versa, it is flagged as a gross inconsistency. More details about how statcheck works can be found in the manual.

statcheck's accuracy

It is important that we know how accurate statcheck is in flagging inconsistencies. We don't want statcheck to mark large numbers of correct results as inconsistent, and, conversely, we also don't want statcheck to wrongly classify results as correct when they are actually inconsistent. We investigated statcheck's accuracy by running it on a set of articles for which inconsistencies were also manually coded.

When we compared statcheck's results with the manual codings, we found two main things. First, statcheck detects roughly 60% of all reported stats. It missed the statistics that were not reported completely according to APA style. Second, statcheck did a very good job in flagging the detected statistics as inconsistencies and gross inconsistencies. We found an overall accuracy of 96.2% to 99.9%, depending on the specific settings. (There has been some debate about this accuracy analysis. A summary of this discussion can be found here.)

Even though statcheck seems to perform well, its classifications are not 100% accurate. But, to be fair, I doubt whether any automated algorithm could achieve this (yet). And again, the comparison with the spellchecker still holds; mine keeps telling me I misspelled my own name, and that it should be "Michelle" (it really shouldn't be).

One major advantage of using statcheck (or any algorithm) for statistical checks is its efficiency. It will take only seconds to flag potential problems in a paper, rather than going through all the reported stats and checking them manually.

An increasing number of researchers seem convinced of statcheck's merits; the R package has been downloaded more than 8,000 times, while the web app has been visited over 23,000 times. Additionally, two flagship psychology journals have started to use statcheck as a standard part of their peer review process. Testimonies on Twitter illustrate the ease and speed with which papers can be checked before they're submitted:

> Just statcheck-ed my first co-authored manuscript. On my phone while brushing my teeth. Great stuff @MicheleNuijten @SachaEpskamp @seanrife!
>
> — Anne Scheel (@annemscheel) October 22, 2016

Automate the error-checking process

More of these “quick and dirty spellchecks” for stats are being developed (e.g. GRIM to spot inconsistencies in means; or *p*-checker to analyse the consistency and other properties of *p*-value), and an increasing number of papers and projects make use of automated scans to retrieve statistics from large numbers of papers (e.g. here, here, here, and here).

In an era where scientists are pressed for time, automated tools such as statcheck can be very helpful. As an author you can make sure you didn’t mistype your key results, and as a peer reviewer you can quickly check if there are obvious problems in the statistics of a paper. Reporting statistics can just as easily go wrong as grammar and spelling; so when you’re typing up a research paper, why not also check your stats?

More information about statcheck can be found at: http://statcheck.io

Note: This article gives the views of the author, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the author

Michèle Nuijten *is an Assistant Professor at Tilburg University, The Netherlands, where she is part of the Meta-Research Center. Her research focuses on the quality of psychological research, including topics such as publication bias, replication, statistical errors, and meta-analysis. Michèle is part of the Executive Board of the Society for the Improvement of Psychological Science (SIPS).*