BayesMed and statcheck

February 28, 2017

Many psychological scientists have blamed the field's replication crisis — which has illuminated the excess of statistically significant findings in the literature — to the fact that most conclusions are based on *p* values (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Critics say *p* values are often wrongly interpreted, can't quantify statistical evidence, and can lead even a null effect to become significant as sample sizes increase (Hoekstra, Finch, Kiers, & Johnson, 2006; Wagenmakers, 2007). There is also evidence that *p* values are often inconsistently reported, which could lead to incorrect conclusions (see, e.g., Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016).

As an attempt to start solving the problems surrounding *p* values, my colleagues and I developed two seemingly very different R packages: BayesMed, a package for a default Bayesian hypothesis test for mediation (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2014; 2015), and statcheck, a package to extract statistical results from articles and recompute *p* values (Epskamp & Nuijten, 2014; Nuijten et al., 2016). In this article, I will explain the rationale behind and use of each of the two R packages, both of which hopefully can improve scientific practice in psychology.

BayesMed

There's a simple way to solve the problems surrounding *p* values: Stop using them. Instead, conclusions can be based on Bayesian statistics. The main principle in Bayesian statistics is that you have a prior belief about an effect, and based on observed data you "update" your prior belief to a posterior belief. This posterior belief is quantified as the probability that your hypothesis is true, given the data. You can also calculate Bayes factors, which (roughly) indicate to what extent one hypothesis is more likely than another.

Using Bayesian statistics has a couple of advantages over using *p* values. For instance, Bayes factors allow you to quantify evidence in favor of or against a hypothesis. If you calculate a Bayes factor for a null hypothesis versus an alternative hypothesis, and you find a Bayes factor of 10, it tells you that the null hypothesis is 10 times more likely than the alternative. If you find a Bayes factor of 1/10, the alternative would be 10 times more likely. Also, a Bayes factor close to 1 tells you that there was insufficient information in the data to draw a conclusion. This method provides a great advantage over using *p* values: With Bayesian statistics, you are able to distinguish between situations in which the null hypothesis is very likely and situations in which your data are ambiguous. A nonsignificant *p* value, on the other hand, can't be used to distinguish between these situations. Furthermore, in the Bayesian framework, you are free to collect more data until a clearer story emerges, since Bayes factors eventually converge to the correct decision. In contrast, *p* values converge to significance when the sample size increases — regardless of the true effect — which leads to an increased rate of false-positive findings.

Although the basic theory underlying Bayesian statistics about updating your beliefs based on data can be quite intuitive, the implementation is often complicated, especially for applied researchers with little

mathematical background. To make Bayesian statistics more accessible, we developed BayesMed: an R package that performs default Bayesian hypothesis tests for correlation, partial correlation, and mediation (Nuijten et al., 2015; Nuijten et al., 2014).

BayesMed doesn't require advanced programming skills. For instance, to test whether the effect of *x* on Y is mediated by M, you need only one line of code: *"jzs_med(independent=X, dependent=Y, mediator=M)."* The "*jzs*" in "*jzs_med*" stands for the default, uninformative "Jeffreys-Zellner-Siow" prior that is used in the calculations. There are similar functions to test for correlation (*jzs_cor*) and partial correlation (*jzs_partcor*). These functions return posterior probabilities and Bayes factors for each of the relations between the independent, dependent, and mediator variable and an overall Bayes factor for mediation (for details, see Nuijten et al., 2015).

We chose to focus on tests for correlation, partial correlation, and mediation because these are among the most common tests used in psychology. For Bayesian alternatives for additional tests, such as *t* tests or analyses of variance, you can use the newly developed (and still developing) software JASP (JASP Team, 2016). JASP offers an easy-to-use "Bayesian SPSS" for common statistical tests. There are plans to incorporate BayesMed's code into JASP, as well.

Software packages such as BayesMed and JASP offer a simple alternative to using *p* values and can hopefully help with redirecting the focus from significant *p* values toward strength of evidence.

The BayesMed R package can be downloaded here.

statcheck

Regardless of whether you agree that Bayesian statistics should be preferred over frequentist statistics, most results in psychology are still based on *p* values, so it is important that these results are at least correctly calculated and reported. However, there is evidence that as many as half of published psychology articles contain at least one result in which the *p* value does not match the reported test statistic and degrees of freedom; additionally, in roughly one out of eight published articles, the reported *p* value leads to a different statistical conclusion than the recomputed *p* value (Bakker & Wicherts, 2014; Caperos & Pardo, 2013; Nuijten et al., 2016; Wicherts, Bakker, & Molenaar, 2011). These inconsistent results can lead to wrong substantive conclusions and affect meta-analyses.

The calculations needed to check the consistency of a result are quite straightforward. However, searching articles by hand to extract statistical results and then recomputing all *p* values is time consuming and error prone. In order to solve this, we developed the R package statcheck (Epskamp & Nuijten, 2014). Statcheck automatically extracts statistical results from articles and recomputes the *p* values. At the moment, statcheck recognizes *t*, *F*, ?2, *r* (correlation), and *Z* tests that are reported in APA style, and when recomputing *p* values it takes into account rounding of the reported test statistic and one-tailed testing.[1]

One of the main advantages of statcheck is that researchers can easily use it to check their manuscript for accidental inconsistencies before submitting to a journal. Besides detecting statistical inconsistencies, statcheck also offers an easy tool to quickly extract published statistics for other analyses. For instance,

with statcheck data, you can estimate the power in a selection of literature or determine effect-size distributions or *p*-value distributions (see, e.g., Hartgerink, Van Aert, Nuijten, Wicherts, & Van Assen, 2016).

There is a detailed manual for statcheck available with instructions on its installation and use. Researchers who are unfamiliar with R can use the new Web app, which includes statcheck's basic functions.

Statcheck can also be used in the review process, both prepublication (see details about the pilot test with statcheck at *Psychological Science* and postpublication (Hartgerink, 2016).

All statcheck data from our 2016 article (Nuijten et al., 2016) are freely available online.

Michèle B. Nuijten will speak at the 2017 APS Annual Convention, May 25–28, 2017, in Boston, Massachusetts.

[1] Note that when one of the three components of an NHST result (test statistic, degrees of freedom, or *p* value) is adjusted to correct for multiple testing, post hoc testing, or violations of assumptions, the result becomes internally inconsistent and statcheck will flag it as such.

References

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of research. *PLoS One, 9*, e103360. doi:10.1371/journal.pone.0103360

Caperos, J. M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema*, *25*, 408–414.

Epskamp, S., & Nuijten, M. B. (2014). statcheck: Extract statistics from articles and recompute p values. R package version 1.0.0. Computer software. Retrieved from http://CRAN.R-project.org/package=statcheck

Hartgerink, C. H. J. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results. *Preprints*, Article 2016080191. doi:10.20944/*preprint*s201608.0191.v1

Hartgerink, C. H. J., Van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & Van Assen, M. A. L. M. (2016). Distributions of *p* values smaller than .05 in psychology: What is going on? *Peerj, 4*, e1935. doi:10.7717/peerj.1935

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review*, *13*, 1033–1037.

JASP Team. (2016). JASP (Version 0.8.0.0). Computer software.

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016).

The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*, 1205–1226. doi:10.3758/s13428-015-0664-2

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (2015). A default Bayesian hypothesis test for mediation. *Behavior Research Methods, 47*, 85–97. doi:10.3758/s13428-014-0470-2

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (2014). BayesMed: Default Bayesian hypothesis tests for correlation, partial correlation, and mediation. R package version 1.0.1. Computer software. Retrieved from https://cran.r-project.org/web/packages/BayesMed/index.html.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779–804. doi:10.3758/BF03194105

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426–432.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One, 6*, e26828. doi:10.1371/journal.pone.0026828