Fewer statistical tests…or better ones?

May, Keith A.; Vincent, Benjamin

Published in:
Perception

DOI:
10.1177/0301006616677909

Publication date:
2017

Document Version
Peer reviewed version

Fewer statistical tests... or better ones?

Keith A. May[1] and Benjamin T. Vincent[2]

1. Department of Psychology, University of Essex
2. School of Psychology, University of Dundee
Address for correspondence: keith.may@essex.ac.uk

In a recent *Perception* editorial, Eli Brenner highlighted the serious problem that many significant results in psychology are not replicable (Brenner, 2016). Brenner identifies a plausible reason for this: that researchers often carry out significance tests on effects that they had not planned to examine in the first place. His solution is to avoid doing any statistical tests that you didn't originally plan to do. We argue that a better solution is to abandon traditional frequentist null-hypothesis significance testing (NHST), and instead use Bayesian data analysis, which can be used to examine unexpected effects without violating the assumptions on which the calculations are based. Bayesians need not worry about polluting the scientific record with false postives just because their question arose after data collection.

Brenner is absolutely correct that, strictly speaking, we should not carry out a traditional test of a null hypothesis unless we had planned to do that test in the first place: the *p*-value is valid only if we stick to exactly what we had planned to do. If we think of an analysis after looking at the data, we can't compute a meaningful *p*-value. This is because the *p*-value is worked out by considering a hypothetical infinite sequence of repetitions of the experiment *and analysis*, and if we allow ourselves to change the experiment or analysis in unpredictable, non-predefined ways, then we can't define what we mean by a repetition, so we can't calculate the *p*-value. This is a major problem, because very often in psychology, we obtain unexpected patterns of results that are potentially very revealing about what is going on. If we analyse these patterns by applying NHST as if we were going to do that all along, we obtain an essentially meaningless *p*-value; this *p*-value can greatly overestimate the true statistical significance of the result because the only reason that we are doing the test is that the data suggested that the difference might be significant. This is why we obtain many false positives.

Brenner's suggested solution is to avoid statistical analysis of any effects that we had not planned to test for. To analyse such unexpected effects using traditional NHST, we should throw away the data and repeat the experiment, this time with the intention of doing the new analysis. But this approach, while technically correct, is hugely wasteful of time and money, and could lead us to ignore potentially exciting discoveries. The data are telling us something, and to refuse to analyse an unexpected effect because our statistical tools can't handle it is like refusing a bowl of soup because there are only forks on the table; a better solution is to find a spoon. In the case of data analysis, the better solution is to use Bayesian methods. These can be applied to any data at any time, regardless of what the experimenter's original intentions were (Berger & Berry, 1988).

The ability to apply a Bayesian analysis to study unexpected effects is a major advantage of Bayesian statistics. But there is another advantage of Bayesian methods which stems from a deep and often-ignored problem with null-hypothesis significance

testing: even if we stick to exactly what we planned to do, and calculate a valid *p*-value, we haven't answered the question we really wanted to ask. The *p*-value gives the probability of obtaining a result as extreme (or more) as the one we did obtain, given that the null hypothesis is true. But that's not what we want to know – we want to know the probability of the null hypothesis given the data, which can be calculated using Bayesian methods. Berger and Berry (1988) give a nice example of a hypothetical experiment and data in which the null hypothesis is rejected at the .05 level, but the probability of the null hypothesis given the data is actually quite high. It is not that NHST gives the wrong answer – it gives the right answer to the wrong question. Bayesian analysis gives the right answer to the right question.

Until relatively recently, uptake of Bayesian data analysis was hindered by the lack of accessible textbooks and the difficulty of calculating the probability distributions. But the recent emergence of introductory texts and free software packages has removed many of the impediments to rational scientific inference. For those raised in the frequentist tradition it can be tricky to get the 'Zen' of the Bayesian approach, but Dienes (2008) provides an engaging introduction of various paradigmatic approaches to inference. The computational implementation of Bayesian methods used to have a high barrier to entry, but software packages such as JAGS (Plummer, 2003) and STAN (Carpenter, Gelman, Hoffman, Lee, & Goodrich, 2016) can now do the heavy lifting. Kruschke (2015) provides a very thorough introduction to the practical aspects of Bayesian data analysis methods using these packages. High-level software packages are now emerging, such as BRMS (https://github.com/paul-buerkner/brms) and RStanArm (https://github.com/stan-dev/rstanarm) allowing users to run common statistical tests with minimal effort in the R statistics environment, for example. At the most accessible end of the spectrum, we have JASP (https://jasp-stats.org), which provides a friendly point-and-click interface to run both frequentist and Bayesian versions of common statistical tests.

We may be on the threshold of a revolution in the way that data are analysed (Brooks, 2003). Indeed, the journal *Basic and Applied Social Psychology* recently banned the use of null hypothesis significance testing, and will no longer publish a paper that reports a *p*-value (Trafimow & Marks, 2015); but they didn't ban Bayesian methods. Bayesian methods are more intuitively appealing; indeed, although most quoted confidence limits are calculated using traditional frequentist methods, a recent survey found that a majority of researchers interpreted confidence intervals as Bayes credible intervals (Hoekstra, Morey, & Rouder, 2014; see the results for question 4 of their questionnaire), which suggests that Bayes credible intervals are what most people really want to calculate anyway. Given the recent proliferation of introductory texts on Bayesian data analysis, and of free software to do this, we would urge researchers to let go of traditional frequentist methods and embrace Bayesian methods, which will almost certainly become the dominant statistical paradigm in the future.

Reference List

Berger, J. O. & Berry, D. A. (1988). Statistical Analysis and the Illusion of Objectivity. *American Scientist, 76*, 159-165.

Brenner, E. (2016). Why we need to do fewer statistical tests. *Perception, 45*, 489-491.

Brooks, S. P. (2003). Bayesian computation: A statistical revolution. *Philosophical Transactions of the Royal Society of London A, 361*, 2681-2697.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., & Goodrich, B. Stan: A probabilistic programming language. *Journal of Statistical Software,* (in press).

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Basingstoke: Palgrave Macmillan.

Hoekstra, R., Morey, R. D., & Rouder, J. N. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*, 1157-1164.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan (2nd ed)*. Waltham, MA: Academic Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (pp. 20-22).

Trafimow, D. & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1-2.