Psychon Bull Rev (2014) 21:309–311
DOI 10.3758/s13423-014-0607-4

Reply to Rouder (2014): Good frequentist properties raise confidence

Adam N. Sanborn · Thomas T. Hills ·
Michael R. Dougherty · Rick P. Thomas ·
Erica C. Yu · Amber M. Sprenger

Published online: 11 March 2014

Abstract Established psychological results have been called into question by demonstrations that statistical significance is easy to achieve, even in the absence of an effect. One often-warned-against practice, choosing when to stop the experiment on the basis of the results, is guaranteed to produce significant results. In response to these demonstrations, Bayes factors have been proposed as an antidote to this practice, because they are invariant with respect to how an experiment was stopped. Should researchers only care about the resulting Bayes factor, without concern for how it was produced? Yu, Sprenger, Thomas, and Dougherty (2014) and Sanborn and Hills (2014) demonstrated that Bayes factors are sometimes strongly influenced by the stopping rules used. However, Rouder (2014) has provided a compelling demonstration that despite this influence, the evidence supplied by Bayes factors remains correct. Here we address why the ability to influence Bayes factors should still matter to researchers, despite the correctness of the evidence. We argue that good frequentist properties mean that results will more often agree with researchers' statistical intuitions, and good frequentist properties control the number of studies that will later be refuted. Both help raise confidence in psychological results.

Keywords Bayesian statistics · Statistical inference · Hypothesis testing

A. N. Sanborn (✉) · T. T. Hills
Department of Psychology, University of Warwick, Coventry CV4 7AL, UK
e-mail: A.N.Sanborn@warwick.ac.uk

M. R. Dougherty · E. C. Yu · A. M. Sprenger
Department of Psychology, University of Maryland, College Park, MD, USA

R. P. Thomas
Department of Psychology, University of Oklahoma, Norman, OK, USA

The recent crisis of confidence in psychology has come about because articles that make exciting claims with apparently solid evidence have been refuted by later work. Although the probability of this happening should be very low, researchers have shown that this can happen more often than standard null-hypothesis significance testing would suggest (Simmons, Nelson, & Simonsohn, 2011). One major concern is researchers who engage in optional stopping of experiments, despite using standard frequentist analyses that assume a fixed number of trials. This can lead to a powerful misrepresentation of the data: With enough time and effort, optional stopping is guaranteed to produce a significant result, even where no effect exists (Armitage, McPherson, & Rowe, 1969).

Though it is generally accepted that optional stopping should be avoided, researchers still engage in this behavior. The prevalence of optional stopping has been highlighted by surveys and experimental evidence (John, Loewenstein, & Prelec, 2012; Yu, Sprenger, Thomas, & Dougherty, 2014). In particular, Yu et al. experimentally demonstrated that professional researchers stop their experiments in order to maximize their chances of obtaining a significant result, or stop early if an experiment does not seem to be going well. The proportion of researchers in Yu et al.'s study who engaged in optional stopping was alarming, especially given the fact that the data for that experiment were collected *after* Simmons et al.'s (2011) highly visible article on researcher degrees of freedom was published.

Bayes factors have been proposed as a solution to this problem, because their interpretation is correct and unchanged for all of the stopping rules that experimenters appear to use (Berger & Wolpert, 1988). This invariance was acknowledged in Yu et al. (2014) and Sanborn and Hills (2014), and has been demonstrated by the simulations of Rouder (2014). We found that Rouder's demonstration, in particular, makes a convincing case that the interpretation of Bayes factors is correct, regardless of the stopping rule. Thus, the assertion made by Yu et al. that Bayes factors are not interpretable if data are collected under an optional stopping rule is clearly an overstatement, since Rouder showed that the Bayes factors retain their interpretability when viewed as a comparison between models.

Although the interpretation of a Bayes factor is correct, no matter the stopping rule used, stopping rules can still influence the data collected, and therefore the Bayes factors produced. Yu et al. (2014), Sanborn and Hills (2014), and Rouder (2014) all showed in simulations that optional stopping influences the resulting Bayes factors. For many situations, the influence is small; for example, it is difficult to produce convincing Bayes factors that support the alternative when the null hypothesis is true (Kerridge, 1963; Smith, 1953). However, Sanborn and Hills showed that in other situations, such as when data are generated by a mixture of two hypotheses or when attempting to find evidence for a null effect, it is much easier to find convincing evidence in favor of either hypothesis, especially when using optional stopping.

The key question, then, is: Does being able to influence Bayes factors through optional stopping matter? Rouder (2014) argued that this is not a problem: Because the Bayes factor always provides a correct interpretation of the evidence obtained, researchers do not have to worry. Despite this, we believe that if researchers wish to use Bayes factors to raise the confidence in psychological results, researchers should care about *how often* particular Bayes factors can be achieved, and not just their value.

A simple reason to care about the frequentist properties of Bayesian statistics is that researchers, inside psychology and out, are more familiar with frequentist statistics. As a result, when frequentist and Bayesian strands of evidence disagree, basing conclusions purely on Bayesian evidence will be unconvincing. For example, researchers trained primarily on frequentist statistics are likely to discount findings in which the Bayes factor indicates evidence for the null hypothesis, yet the effect is significant.[1] Being able to demonstrate that your results also satisfy the common frequentist intuitions of researchers goes a long way toward instilling confidence in your results.

A more complex reason is that frequentist properties are important for many common ways of evaluating scientific research, both for publication and for professional advancement. Certainly articles that present surprising conclusions supported by convincing evidence have the best chance for publication. Producing articles, especially those that appear in high-impact journals, is necessary for academic rewards. Article counts are important, and having more of them in high-impact journals is better. The chances of finding strong evidence, which often govern whether an article is publishable, and thus how many articles are produced, are frequentist properties. Frequentist properties, such as how often a finding has been replicated, are also more easily interpreted by journalists and the general public, who mostly lack training in statistics or probabilistic reasoning.[2]

Yu et al. (2014) showed that professional researchers, motivated by costs and rewards, use optional stopping to find publishable evidence. It is no leap to suppose that if Bayes factors and optional stopping were generally considered safe, then researchers would find the easiest ways to produce evidence under the new rules. Situations in which it is relatively easy to produce convincing evidence in either direction would be particularly attractive. If large numbers of articles of this type were produced, many of which were later refuted, it would place psychological research back into its current crisis of confidence. Thus, paying attention to the frequentist properties of Bayesian tests to minimize the number of these articles seems like the safest route to raising confidence.

We agree that Bayes factors provide a valuable and coherent approach to statistical inference, and we believe that they should be more widely adopted and used, as we have detailed elsewhere (e.g., Sprenger et al., 2013; Tidwell, Dougherty, Chrabaszcz, Thomas, & Mendoza, 2013). If the academic culture were to change, then one could envision a world in which Bayesian statistics would be used exclusively, without regard for frequentist concerns. Unfortunately, the field has a long way to go before that will be the case, as is evidenced by the fact that only 1 of the 314 respondents in Yu et al. (2014) reported using Bayesian methods. Thus, as things stand, it is important to ensure that statistical tests have good frequentist properties, so that they satisfy the intuitions of consumers of our science, whether they be journalists or academics. This could be done in many different ways, some of which are outlined in Sanborn and Hills (2014). Therefore, we echo Rosenbaum and Rubin's (1984) statement that "Bayesians as well as frequentists need to attend carefully to the procedures used to collect data" (p. 108).

[1] This can happen in various situations, such as when researchers are very uncertain about the size of the effect and have yet to collect sufficient data to overcome the uncertainty in their prior (Lindley, 1957). It can also happen after a great many data have been collected: If significance is barely achieved, the Bayes factor, using any reasonable continuous prior, will strongly favor the null hypothesis (Wagenmakers & Grünwald, 2006).

[2] One example of frequentist properties being used for communicating findings recently appeared in a book on cognitive training by Dan Hurley (2014). In making the argument that cognitive training can improve cognitive abilities, Hurley stated, "I am aware of seventy-five randomized trials, published in peer-reviewed scientific journals, that have found a significant benefit to cognitive training of various sorts, and a grand total of four that have found no such benefit" (p. 152). Disregarding the question of whether all 75 articles really did find meaningful improvements, the implications of the statement are obvious.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A, 132,* 235–244.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*. Hayward: Institute of Mathematical Statistics.

Hurley, D. (2014). *Smarter: The new science of building brain power.* New York: Hudson Street Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. doi:10.1177/0956797611430953

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *Annals of Mathematical Statistics, 34,* 1109–1110.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44,* 187–192.

Rosenbaum, P. R., & Rubin, D. B. (1984). Sensitivity of Bayes inference with data-dependent stopping rules. *American Statistician, 38,* 106–109.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review.*

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review.* doi:10.3758/s13423-013-0518-9

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Smith, C. (1953). The detection of linkage in human genetics. *Journal of the Royal Statistical Society: Series B, 15,* 153–192.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Weems, S. A., . . . Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence*, 41, 638–663. doi:10.1016/j.intell.2013.07.013

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2013). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin and Review.* doi:10.3758/s13423-013-0560-7

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science, 17,* 641–642. doi:10.1111/j.1467-9280.2006.01757.x

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review.* doi:10.3758/s13423-013-0495-z