Reply to Rouder (2014): Good Frequentist Properties Raise Confidence

Adam N. Sanborn[1], Thomas T. Hills[1], Michael R. Dougherty[2], Rick P. Thomas[3], Erica C. Yu[2], Amber M. Sprenger[2]

[1]Department of Psychology, University of Warwick, Coventry CV4 7AL, UK

[2]Department of Psychology, University of Maryland, College Park, MD 20742

[3]Department of Psychology, University of Oklahoma, Norman OK, 73019

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 based on 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 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 (in press) and Sanborn and Hills (in press) demonstrated that Bayes factors are sometimes strongly influenced by the stopping rules used. However, Rouder (in press) 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 papers that will later be refuted. Both help raise confidence in psychological results.

Keywords: Bayesian statistics, statistical inference, hypothesis testing

Reply to Rouder (2014): Good Frequentist Properties Raise Confidence

The recent crisis of confidence in psychology has come about because papers that make exciting claims with apparently solid evidence have been refuted by later work. While 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 engaging 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 find a significant result 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, in press). In particular Yu et al. (in press) experimentally demonstrated that professional researchers stop their experiments in order to maximize their chance 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. who engaged in optional stopping was alarming, especially given the fact that the data for that experiment was collected *after* Simmons et al.'s (2011) highly visible paper 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 the stopping rules experimenters appear to use (Berger & Wolpert, 1988). This invariance was acknowledged in Yu et al. (in press) and Sanborn and Hills (in press) and demonstrated by the simulations of Rouder (in press). 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 in Yu et al. (in press) that Bayes factors are not interpretable if data are collected under an optional stopping rule is clearly an overstatement, as Rouder (in press) shows that the Bayes factors retain their interpretability when viewed as a comparison between models.

While 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. (in press), Sanborn and Hills (in press), and Rouder (in press) 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 (in press) showed that there are other situations, such as when the data are generated by a mixture of two hypotheses or when attempting to find evidence for a null effect, in which it is much easier to find convincing evidence in favor of either hypothesis, especially when using optional stopping.

The key question is then: Does being able to influence Bayes factors through optional stopping matter? Rouder (in press) argues that it 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 evidence disagrees, then basing conclusions purely on Bayesian evidence will be unconvincing. For example, researchers trained primarily on frequentist statistics are likely to discount findings where 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 towards 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 papers that present surprising conclusions supported by convincing evidence have the best chance for publication. Producing papers, especially those that appear in high impact journals, is necessary for academic rewards. Paper counts are important: having more papers in high impact journals is better. The chances of finding strong evidence, which often governs whether a paper is publishable and thus how many papers 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. (in press) 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. The situations in which it is relatively easy to produce convincing evidence in either direction would be particularly attractive. If large numbers of papers of this type were produced, many which later are 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 papers 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 they should be more widely adopted and used, as we've detailed elsewhere (e.g., Sprenger et al., 2013; Tidwell, Dougherty, Chrabaszcz, Thomas, & Mendoza, in press). If the academic culture were to change, then one could envision a world in which Bayesian statistics are used exclusively without regards for frequentist concerns. Unfortunately, the field has a long ways to go before that will be the case, as evidenced by the fact that only 1 of the 314 respondents in Yu et al. reported using Bayesian methods. Thus, as it stands 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. There are many different ways that this could be done,

[1] This can happen in various situations. It can happen 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 lot of 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 where frequentist properties were used for communicating findings recently appeared in a book on cognitive training by Dan Hurley. In making the argue that cognitive training can improve cognitive abilities, Hurley states "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." Disregarding the question over whether all 75 articles really did find meaningful improvement, the implications of the statement are obvious.

some of which are outlined in Sanborn and Hills (in press). 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)."

References

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

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

Hurley, D. (2014). Smarter: The new science of building brain power. Huxley

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.

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

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44(1/2),* 187-192.

Rouder, J. N. (in Press). Optional Stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*.

Rosenbaum, P. R., & Rubin, D. B. (1984). Sensitivity of Bayes inference with data-dependent stopping rules. *The American Statistician*, *38*(2), 106-109.

Sanborn, A. N. & Hills, T. T. (in press). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*.

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*(11), 1359-1366.

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., Chrabaszcz, J.S., Smith, V., Bobb, S., Bunting, M. F., & Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence, 41,* 638–663. doi.org/10.1016/j.intell.2013.07.013

Tidwell, J., Dougherty, M. R., Chrabaszcz, J., Thomas, R. P., & Mendoza, J. (in press). What counts as evidence for working memory training? *Psychonomic Bulletin and Review*.

Wagenmakers, E. J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science*,*17*(7), 641-642.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (in press). When decision heuristics and science collide. *Psychonomic Bulletin & Review*.