Comment: A simple alternative to p-values

Daniel J. Benjamin (Univ. of Southern California) and James O. Berger (Duke Univ.)

The ASA statement is excellent, and we completely endorse it. Unfortunately, there is a long history of such statements appearing and having only transient impact, because an alternative to p-values is not specified. There have been many proposed alternatives that are much more readily interpretable than p-values, but none has gained widespread acceptance. We believe that p-value alternatives have failed to garner general support because they have either or both of two shortcomings:

1. Being more complicated than p-values.

2. Not being acceptable to both frequentist and Bayesian schools of statistical thought.

Recently, in Bayarri et al. (2016), we have proposed an alternative that overcomes both of these hurdles.

Our proposal is a follow-up to suggestions (cf., Wellcome Trust Case Control Consortium (2007)) that a researcher should report the *pre-experimental odds of correct rejection to incorrect rejection* of the null hypothesis. If, for example, these odds are 10 to 1, then a formal rejection of the null hypothesis is 10 times more likely to be a correct rejection than an incorrect rejection. But the pre-experimental odds depend partly on the researchers *prior odds* of the alternative hypothesis to the null hypothesis, which makes them non-frequentist.

The pre-experimental odds can be decomposed into the product of the prior odds and a frequentist component, which is determined by the experimental design and planned analysis. We can then focus attention on this frequentist component, which we call the *rejection ratio*. It is the probability of rejection when the alternative hypothesis is true, divided by the probability of rejection when the null hypothesis is true, i.e., the ratio of the power of the experiment to the Type I error of the experiment. The rejection ratio has a straightforward interpretation as quantifying the strength of evidence about the alternative hypothesis relative to the null hypothesis conveyed by the experimental result being statistically significant.

While the rejection ratio is an excellent summary of the quality of the experiment in terms of hypothesis testing, it would not work as an alternative to p-values because it suffers from the two drawbacks listed above. First, it is more complicated than a p-value because it requires power

calculations. Second, although justified to unconditional frequentists, it is unsatisfactory to many (including Bayesians) because it is not data-dependent; if the rejection ratio were 10 based on a Type I error of $\alpha = 0.05$, then 10 would be reported regardless of whether the data yields a p-value of 0.05, right at the boundary of the rejection region, or a p-value of 0.000001, surely indicating much stronger evidence against the null hypothesis than $p = 0.05$. To a Bayesian, the correct data-dependent measure of the evidence in favor of the alternative hypothesis relative to the null hypothesis is the Bayes factor, given by

$$B = \frac{\text{average likelihood of the observed data under the alternative hypothesis}}{\text{likelihood of the observed data under the null hypothesis}}.$$

It would seem that we are now at a Bayesian/frequentist impasse, but this is not so - at least, it is not so for many common situations such as testing a null hypothesis of zero effect versus a two-sided alternative hypothesis of non-zero effect.[1] Indeed, for such situations, we show in Bayarri et al. (2016) that B, while data dependent, is a fully frequentist measure because its frequentist expectation under the null hypothesis precisely equals the frequentist rejection ratio. Bayesians and frequentists should thus unite in promoting B as an easily interpretable alternative to p-values, overcoming problem #2 above.

Unfortunately, B still suffers from problem #1 because the 'average likelihood' in the numerator of B needs to be computed using some assumed prior distribution for the alternative hypothesis. There is, however, a simple upper bound on B (Vovk (1993)) that holds under quite general

[1] More generally, our recommendation here applies to any situation of *precise hypothesis testing*, by which we mean that the null hypothesis is a lower dimensional subspace of the alternative hypothesis, as in testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The major problems with p-values are muted if the hypotheses are, say, $H_0 : \theta < 0$ versus $H_1 : \theta > 0$, which are of the same dimension. As an example, suppose θ denotes the difference in mean treatment effects for cancer treatments A and B:

- Scenario 1: Treatment A = standard chemotherapy and Treatment B = standard chemotherapy + steroids. This is a scenario of precise hypothesis testing, because steroids could be essentially ineffective against cancer, so that θ could quite plausibly be essentially zero.
- Scenario 2: Treatment A = standard chemotherapy and Treatment B = a new radiation therapy. In this case there is no reason to think that θ could be zero, and it would be more appropriate to test $H_0 : \theta < 0$ versus $H_1 : \theta > 0$.

See Berger and Mortera (1999) for discussion of these issues. Moreover, in precise hypothesis testing situations that are one-sided, such as $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, $\overline{B}$ is no longer strictly an upper bound for B (although the deviations tend to be minor; Sellke (1977)).

conditions (Sellke et al. (2001)), namely

$$B \leq \overline{B} \equiv \frac{1}{-e\ p \log p}\,. \tag{1}$$

The Bayes factor bound $\overline{B}$ is the largest possible B over any (reasonable) choice of the prior distribution for the alternative hypothesis. Like B, $\overline{B}$ overcomes problem #2: it is justifiable to both Bayesians and frequentists. However, $\overline{B}$ additionally surmounts problem #1 because it is a simple function of the p-value.

The following table shows the value of $\overline{B}$ for a wide range of p-values.

p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001
$\frac{1}{-ep\log(p)}$	1.60	2.44	8.13	13.9	52.9	400	3226

An important implication of these calculations is that results that just reach conventional levels of significance do not actually provide very strong evidence against the null hypothesis. For example, a p-value of 0.05 corresponds to a Bayes factor of at most 2.44 : 1. That is, the data imply odds in favor of the alternative hypothesis relative to the null hypothesis of at most 2.44 to 1. A p-value of 0.01 - often considered 'highly significant' – corresponds to at most 8.13 to 1 odds, hardly overwhelmingly convincing odds.

Since $\overline{B}$ indicates the strongest potentially justifiable inference from the data, its use would alert researchers when seemingly strong evidence is actually not very compelling. Its use would therefore help prevent researchers from being misled into concluding too much from the statistical significance of a finding. Interestingly, although $\overline{B}$ is only an upper bound on the Bayes factor, we report evidence in Bayarri et al. (2016) that, when calculated from real data from a range of scientific fields, $\overline{B}$ is often not that far from the B implied by a scientifically reasonable alternative hypothesis.

In short, to begin the process of recovering from the misuse of p-values, we propose replacing the p-value by the Bayes factor bound $\overline{B}$. $\overline{B}$ is easily interpretable, has both Bayesian and frequentist justification, and is as simple to calculate as the p-value. Of course, we would encourage everyone to 'bite the bullet' and use the more sophisticated B – again, also a fully frequentist measure – and even to go one step further by multiplying B by the prior odds to obtain the overall odds of the alternative hypothesis to the null hypothesis. But even just the initial, small step of reporting $\overline{B}$ would help researchers avoid some of the most problematic and apparently inevitable misinterpretations that

arise from reliance on the p-value.

Acknowledgments: This research was supported by the National Science Foundation, under grants DMS-1007773, DMS-1407775, and BCS-1521855, and the National Institutes of Health / National Institute on Aging under grant R01-AG042568.

References

Bayarri, M.J., Benjamin, D., Berger, J., and Sellke, T. (2016). Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses. To appear in *J. Mathematical Psychology.*

Berger, J. and Mortera, J. (1999). Default Bayes factors for non-nested hypothesis testing. *J. Amer. Statist. Assoc.*, **94**, 542–554.

Sellke, T., Bayarri, M.J., and Berger, J.O. (2001). Calibration of p Values for Testing Precise Null Hypotheses, *The American Statistician* **55**, 62-71.

Sellke, T.M. (2012). On the interpretation of p-values, *Tech. Rep. Department of Statistics, Purdue University.*

Vovk, V.G. (1993). A Logic of Probability, with Application to the Foundations of Statistics. *Journal of the Royal Statistical Society. Series B*, **55**, 317–351

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature* **447**(7145), 661-678.