The p-value of a simple hypothesis is typically much much smaller than the Bayes factor

Jonathan Rougier*

School of Mathematics

University of Bristol

Compiled on October 11, 2016, from file `pvalue4.Rnw`

Abstract

An 'embedding model' for a simple hypothesis justifies the choice of test statistic when computing a p-value. Under this embedding model, the p-value is bounded above by every Bayes factor. The nature of the bound suggests that the p-value is typically much much smaller than any reasonable Bayes factor.

Keywords: Simple hypothesis, P-value, Bayes factor, embedding model

This is a brief contribution to the ongoing discussion about the evidential import of a small p-value (Wasserstein and Lazar, 2016, this Journal). Let $X \in \mathcal{X}$ be a set of observables, and $H_0 : X \sim f_0$ be a simple hypothesis. A 'significance procedure' for H_0 is any statistic $p_0 : \mathcal{X} \to \mathbb{R}$ such that

*School of Mathematics, University Walk, Bristol BS8 1TW, UK; email `j.c.rougier@bristol.ac.uk`.

$p_0(X)$ under H_0 stochastically dominates a uniform distribution. If p_0 is a significance procedure for H_0, then $p_0(x^{\text{obs}})$ is a 'p-value', where x^{obs} are the observations of X. The usual way to construct a significance procedure is to propose a test statistic $t : \mathcal{X} \to \mathbb{R}$. Then

$$p_0(x) := \Pr_0\{t(X) \geq t(x)\} \tag{1}$$

is a significance procedure according to the Probability Integral Transform, where $\Pr_0$ is the probability under H_0. For more on these definitions, see Casella and Berger (2002, sec. 8.3). I find the distinction between a 'procedure' and a 'value', which I took from Morey et al. (2016), to be very useful in practice.

Clearly there are an uncountable number of significance procedures for H_0, one for each choice of t. Presumably most of them are not very informative for the question of interest. Therefore we do the analyst the courtesy of assuming that for her p-value, the test statistic t was carefully chosen to reflect the question of interest. From this viewpoint, we can propose an embedding model for X in which t is an unambiguously good choice for testing H_0 versus 'not H_0', as was originally suggested by David Cox, in Savage et al. (1962, p. 84) and Cox (1977). Cox suggested the exponentially-tilted null model,

$$f(x;\theta) = \frac{f_0(x) \cdot e^{\theta \cdot t(x)}}{M_T(\theta)}, \qquad \theta \geq 0, \tag{2}$$

where M_T is the Moment Generating Function of $t(X)$ under H_0. This model has a Monotone Likelihood Ratio in $t(x)$, and hence the test statistic t is

Uniformly Most Powerful in testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ (e.g. Casella and Berger, 2002, sec. 8.3).

This is a 'sufficient' argument for the embedding model; i.e. were (2) the model, then t would be the analyst's unambiguous choice of test statistic for H_0 versus 'not H_0'. But it is also hard to imagine a simpler way to create an embedding model out of just f_0 and t, and this might be a more practical justification for (2). However, the following argument generalizes from (2) to a large and appropriate class of embedding models, as shown below in Theorem 1.

Consider the Bayes factor for H_0 versus H_1,

$$B_{01}(x) := \frac{f_0(x)}{\int_0^\infty f(x;\theta)\,\pi(\theta)\,\mathrm{d}\theta}, \tag{3}$$

where π is some prior distribution on $\theta \in (0, \infty)$. Adopting the approach originally proposed by L.J. Savage (Edwards et al., 1963, p. 228), the Bayes factor can be bounded below over the set of all possible priors,

$$B_{01}(x) \gg \inf_\pi \frac{f_0(x)}{\int_0^\infty f(x;\theta)\,\pi(\theta)\,\mathrm{d}\theta} = \inf_{\theta>0} \frac{f_0(x)}{f(x;\theta)}, \tag{4}$$

where I have written '$\gg$' for 'much greater than', which seems justified, given that the infimum happens at a prior which is a delta function, and which would never be chosen in practice. But then, from (2),

$$\inf_{\theta>0} \frac{f_0(x)}{f(x;\theta)} = \inf_{\theta>0} e^{-\theta\cdot t(x)}\, M_T(\theta) \gg \Pr_0\{t(X) \geq t(x)\} = p_0(x) \tag{5}$$

according to Chernoff's inequality (e.g. Whittle, 2000, ch. 15). I have

written '$\gg$' again, because Chernoff's inequality is an application of Markov's inequality, which is typically very generous, although there is a *caveat* for IID X's arising from Extreme Value Theory, (e.g. Whittle, 2000, ch. 18). Putting (4) and (5) together,

$$p_0(x) \ll \cdots \ll B_{01}(x), \tag{6}$$

typically, from whence the title of this note.

How small is 'much much smaller'? Consider the canonical statistical model, first analysed in this context by Edwards et al. (1963, p. 228). Let the null model be $X \sim \mathrm{N}(0, \sigma^2)$ for known σ, and let the test statistic be $t(x) = x/\sigma$. Then the embedding model from (2) is $X \sim \mathrm{N}(\theta, \sigma^2)$ and

$$B_{01}(x) \geq \exp\left\{-\tfrac{1}{2}(x/\sigma)^2\right\}. \tag{7}$$

Figure 1 plots the p-value and the lower bound for B_{01}. Also shown are some specific values: the lower bound on $B_{01}(x)$ when $p_0(x) = 0.05$, and the value of $p_0(x)$ corresponding to a lower bound of $B_{01}(x) \geq 10^{-3/2} \approx 0.032$, which is the boundary between 'strong' and 'very strong' evidence against H_0 in the scheme of Harold Jeffreys (Jeffreys, 1961, Appendix B). In this example, a p-value at the conventional threshold of 0.05 would correspond to a lower bound for the Bayes factor within the embedding model of 0.259, or more than five times larger. From the other direction, the *necessary* condition for satisfying Jeffreys's boundary is $p_0(x) < 0.004$. All in all, the conventional threshold for rejecting H_0 of $p_0(x) < 0.05$ seems positively reckless, with a

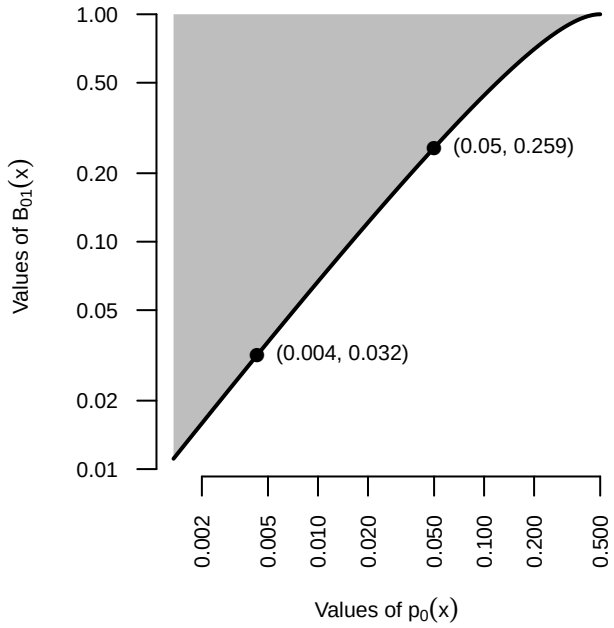

Figure 1: The set of possible values for the p-value and Bayes factor, for the null model $X \sim \mathrm{N}(0, \sigma^2)$ and the test statistic $t(x) = x/\sigma$, under the embedding model (2).

value one tenth the size being more appropriate. Johnson (2013) recommends similar thresholds (0.005 and 0.001), but with different reasoning.

A similar result to (6) holds for a much more general class of embedding models, based on f_0 and t. Here is the exact result.

Theorem 1. *Let f_0 and t be given. Let the set of embedding models be*

$$f(x;\theta) \propto f_0(x) \cdot g\big(t(x),\theta\big), \qquad \theta \geq 0 \tag{8}$$

for some non-negative g for which $g(\cdot\,,\theta)$ is increasing, and $g(\cdot\,,0) = 1$. Then $p_0(x) \leq B_{01}(x)$.

Proof.

$$
\begin{aligned}
p_0(x) &= \Pr_0\{t(X) \geq t(x)\} \\
&= \Pr_0\{g(t(X),\theta) \geq g(t(x),\theta)\} && g(\cdot\,,\theta) \text{ increasing} \\
&\leq \frac{\mathrm{E}_0\{g(t(X),\theta)\}}{g(t(x),\theta)} && g \geq 0, \text{ Markov's inequality} \\
&\leq \inf_{\theta>0} \frac{\mathrm{E}_0\{g(t(X),\theta)\}}{g(t(x),\theta)} && \text{holds for all } \theta \\
&= \inf_{\theta>0} \frac{f_0(x)}{f(x;\theta)} && \text{by (8)} \\
&\leq B_{01}(x) && \text{by (4).} \qquad \square
\end{aligned}
$$

Thus (6) gains additional support from being true over a large class of embedding models. Eq. (8) identifies $t(X)$ as a sufficient statistic for θ in the embedding model, according to the Fisher-Neyman Factorization Theorem (e.g. Casella and Berger, 2002, sec. 6.2); the other conditions on g are trivial. Hence the following *précis* of Theorem 1, which blurs the distinction between the technical and the vernacular meaning of 'sufficient', but which is easy to grasp.

> *If the statistic t is sufficient to assess departures from the null model f_0, then the p-value is typically much much smaller than the Bayes factor. And if t is not sufficient to assess departures from f_0, then it would be foolish to assess f_0 on the basis of the p-value alone.*

In this form, Theorem 1 serves as a complement to Lindley's paradox (Lindley, 1957). In its most primitive form, Lindley's paradox states that if H_0

and H_1 are two simple hypotheses, then it is possible that $B_{01}(x) \geq 1/p_0(x)$ for some $x \in \mathcal{X}$; i.e. the p-value for H_0 may be very small while the Bayes factor for H_0 versus H_1 may be very large. Theorem 1 states that $p_0(x) \leq B_{01}(x)$ for all $x \in \mathcal{X}$, whenever H_1 is chosen in a sensible way. Both results have the same implication, which is to impugn the use of a p-value for 'rejecting' H_0 (Wasserstein and Lazar, 2016, points 3 and 6). Lindley's paradox says that if you use a fixed threshold, you can sometimes be spectacularly wrong. Theorem 1 says that the threshold $p_0(x) < 0.05$ provides little evidence against H_0, in the class of embedding models that are consistent with the test statistic.

References

Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44:533–534.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition.

Cox, D. (1977). The role of significance tests. *Scandinavian Journal of Statistics*, 4(3):49–70. With discussion and rejoinder.

Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242.

Jeffreys, H. (1961). *Theory of Probability*. Oxford, UK: Oxford University Press, 3rd edition.

Johnson, V. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317.

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192. See also Bartlett (1957).

Morey, R., Hoekstra, R., Rouder, J., Lee, M., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bullentin & Review*, 23(1):103–123.

Savage, L. et al. (1962). *The Foundations of Statistical Inference*. Methuen, London, UK.

Wasserstein, R. and Lazar, N. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.

Whittle, P. (2000). *Probability via Expectation*. New York: Springer, 4th edition.