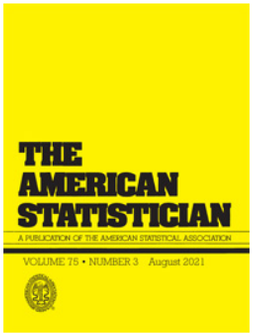
Taylor & Francis
Taylor & Francis Group



The American Statistician

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Null Hypothesis Significance Testing Defended and Calibrated by Bayesian Model Checking

David R. Bickel

To cite this article: David R. Bickel (2021) Null Hypothesis Significance Testing Defended and Calibrated by Bayesian Model Checking, The American Statistician, 75:3, 249-255, DOI: 10.1080/00031305.2019.1699443

To link to this article: https://doi.org/10.1080/00031305.2019.1699443

THE AMERICAN STATISTICIAN
2021, VOL. 75, NO. 3, 249–255: General
https://doi.org/10.1080/00031305.2019.1699443

Taylor & Francis
Taylor & Francis Group

Null Hypothesis Significance Testing Defended and Calibrated by Bayesian Model Checking

David R. Bickel[a,b,c]

[a]Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON, Canada; [b]Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, ON, Canada; [c]Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada

ABSTRACT

Significance testing is often criticized because *p*-values can be low even though posterior probabilities of the null hypothesis are not low according to some Bayesian models. Those models, however, would assign low prior probabilities to the observation that the *p*-value is sufficiently low. That conflict between the models and the data may indicate that the models needs revision. Indeed, if the *p*-value is sufficiently small while the posterior probability according to a model is insufficiently small, then the model will fail a model check. That result leads to a way to calibrate a *p*-value by transforming it into an upper bound on the posterior probability of the null hypothesis (conditional on rejection) for any model that would pass the check. The calibration may be calculated from a prior probability of the null hypothesis and the stringency of the check without more detailed modeling. An upper bound, as opposed to a lower bound, can justify concluding that the null hypothesis has a low posterior probability.

ARTICLE HISTORY
Received March 2019
Accepted November 2019

KEYWORDS
Hypothesis testing; Model checking; Objective Bayes factor; *p*-Value calibration; Relative belief ratio; Reproducibility crisis

1. Introduction

The mounting opposition against null hypothesis significance testing ranges from warnings about misusing it (Wasserstein and Lazar 2016) to its outright banning from journal publication (Trafimow and Marks 2015). The most trenchant criticisms come from supporters of the likelihood principle, especially Bayesians.

Example 1. Consider a sample x of n observations $x_1, x_2, \ldots, x_n$ that are modeled as realized values of the independent random variables $X_1, X_2, \ldots, X_n$ drawn from the normal distribution of mean θ and standard deviation σ. Let $\bar{x}$ and s denote the observed values of the sample mean $\bar{X}$ and its unbiased sample standard error S. To test the null hypothesis $H_0 : \theta = 0$ against the alternative hypothesis $H_1 : \theta \neq 0$, a one-sample t-test could be used. For a large sample size n, the two-sided p-value is approximated by tail areas of a normal distribution:

$$p = \Pr(|Z| \geq |z| \,|\, H_0) \approx 2 \min(\Phi(z), 1 - \Phi(z)) = 2(1 - \Phi(|z|)),$$

where $z = \bar{x}/s$, $Z = \bar{X}/S$, and Φ is the cumulative distribution function of the standard normal distribution. Note that z is the number of sample standard errors of the observed sample mean from 0.

For any $k > 0$, the test yielding the above approximate p value is called a *k-sigma test* if H_0 is rejected whenever $|z| > k$, that is, whenever $p < 2(1 - \Phi(k))$. For example, if $k = 5$, then the significance level is $\alpha = 2(1 - \Phi(5)) \approx 5 \times 10^{-7}$. Since that α is so small, if $p < \alpha$, then "*Either* an exceptionally rare chance has occurred, *or* the [null hypothesis] is not true" (Fisher 1973, p. 42).

Regarding the 5-sigma test associated with the discovery of the Higgs boson, O'Hagan (2012) remarked,

> Five standard deviations, assuming normality, means a p-value of around 0.0000005…. Rather than ad hoc justification of a p-value, it is of course better to do a proper Bayesian analysis…. We know that given enough data it is nearly always possible for a significance test to reject the null hypothesis at arbitrarily low p-values, simply because the parameter will never be exactly equal to its null value. And apparently the LHC has accumulated a very large quantity of data. So could even this extreme p-value be illusory?

In other words, while the p-value could be low because the Higgs boson was discovered, it might instead be low due to systematic measurement error that is not modeled by the point null hypothesis. Even apart from such systematic measurement error, the discovery could well be illusory with very high probability.

An example of such an illusion is that in which a traditional Bayesian analysis with strictly positive prior probability of a sharp null hypothesis such as H_0 yields a very high posterior probability of the null hypothesis despite a low p-value. That can happen not only for the posterior probability $\Pr(H_0 \,|\, X = x)$, where $X = (X_1, X_2, \ldots, X_n)$ and $x = (x_1, x_2, \ldots, x_n)$ (e.g., Bartlett 1957; Lindley 1957), but also, assuming $p \leq \alpha$, for the posterior probability $\Pr(H_0 \,|\, P \leq \alpha)$, where P is the unknown p-value modeled as a random variable. To see how that is possible, invoke Bayes' theorem:

$$\frac{\Pr(H_0 \,|\, P \leq \alpha)}{\Pr(H_0)} = \frac{\Pr(P \leq \alpha \,|\, H_0)}{\Pr(P \leq \alpha)} = \frac{\alpha}{\Pr(P \leq \alpha)}. \quad (1)$$

CONTACT David R. Bickel dbickel@uottawa.ca Department of Biochemistry, Microbiology and Immunology, University of Ottawa, 451 Smyth Road, Ottawa, ON K1H 8M5, Canada.

Thus, even if α is small, the posterior probability of H_0 can be high relative to its prior probability, provided that $\Pr(P \leq \alpha)$ is not much larger than α. In that case, however, $\Pr(P \leq \alpha)$ would be small even though $p \leq \alpha$ was observed. Recycling the words of Fisher, "*Either* an exceptionally rare chance has occurred, *or*" the Bayesian model behind the probability function Pr "is not true." The influence of Bayesian modeling on $\Pr(P \leq \alpha)$ is clarified by

$$\begin{aligned}\Pr(P \leq \alpha) &= \Pr(H_0)\Pr(P \leq \alpha \mid H_0) + \Pr(H_1)\Pr(P \leq \alpha \mid H_1) \\ &= \Pr(H_0)\,\alpha + (1 - \Pr(H_0)) \int \Pr(P \leq \alpha \mid \theta)\,\pi_1(\theta)\,d\theta, \quad (2)\end{aligned}$$

where $\pi_1(\theta)$ is the prior probability density of θ conditional on H_1. The Bayesian model that specifies the prior probability $\Pr(H_0)$ and the prior probability density function π_1 fails to predict the observation that $p \leq \alpha$ whenever $\Pr(P \leq \alpha)$ is low. As we have seen, that occurs whenever α is low while $\Pr(H_0 \mid P \leq \alpha)$ is relatively high.

The Fisherian either-or reasoning motivates using a Bayesian p-value to check priors and other aspects of a Bayesian model in the same way that a p-value, such as Example 1's $2(1 - \Phi(|z|))$, checks a null hypothesis (e.g., Micheas and Dey 2003). In the notation of Example 1, one such Bayesian p-value is the prior predictive p-value $\Pr(|Z| \geq |z|)$, which checks the Pr-function's Bayesian model in the same way that $p = \Pr(|Z| \geq |z| \mid H_0)$ tests H_0. That prior predictive p-value is $\Pr(P \leq p)$, which applies in general, not just for Example 1. If $\Pr(P \leq p)$ is below some threshold γ, then the Bayesian model is said to *fail* the check at level γ; otherwise, it *passes* the check.

Example 2. Consider a normal random variable X of unit mean and unknown standard deviation θ. The null hypothesis is $H_0 : \theta = 1$, and the observed value of X is $x = 3$, which implies that the two-sided p-value is $p = 0.003$, just under $\alpha = 0.005$, the significance level recommended by Benjamin et al. (2017). The alternative hypothesis is $H_1 : \theta = 1.1$, leading to

$$\begin{aligned}\Pr(P \leq 0.005 \mid H_1) &= \Pr\left(|Z| \geq \Phi^{-1}(1 - 0.005/2) \mid H_1\right) \\ &= 2\left(1 - \Phi\left(\frac{\Phi^{-1}(1 - 0.005/2)}{1.1}\right)\right) \\ &\approx 2\left(1 - \Phi\left(\frac{2.8}{1.1}\right)\right) \approx 0.01.\end{aligned}$$

If the prior odds $\Pr(H_0)/\Pr(H_1)$ is 10, an empirically supported default (Benjamin et al. 2017), then $\Pr(H_0) = 10/11$, and Equations (1) and (2) tell us that

$$\begin{aligned}\Pr(H_0 \mid P \leq \alpha) &= \Pr(H_0 \mid P \leq 0.005) \\ &= \frac{(10/11)(0.005)}{(10/11)(0.005) + (1/11)\Pr(P \leq 0.005 \mid H_1)} \quad (3) \\ &\approx \frac{0.005}{0.005 + (0.1)(0.01)} = \frac{0.005}{0.006} = \frac{5}{6}.\end{aligned}$$

In short, $\Pr(H_0 \mid P \leq 0.005)$ is high even though $p \leq 0.005$. However, that is only because $\Pr(P \leq 0.005)$, given in the denominator of Equation (3), is very low: $\Pr(P \leq 0.005) = 0.006$. *Either* an event of probability 0.6% occurred *or* the Bayesian model encoded in the function Pr is "not true," perhaps because $\theta = 1.1$ is inadequate as the alternative hypothesis. Accordingly, the prior predictive p-value is even smaller:

$$\begin{aligned}\Pr(P \leq 0.003) &= (10/11)(0.005) + (1/11)\Pr(P \leq 0.003 \mid H_1) \\ &= (10/11)(0.005) + (1/11)\Pr(P \leq 0.003 \mid H_1) \\ &\approx (10/11)(0.005) + (1/11)(0.007) \approx 0.005,\end{aligned}$$

small enough that the Bayesian model would fail the check for any $\gamma \geq 0.006$.

Notice that $p \leq \alpha$ implies that $\Pr(P \leq p) \leq \Pr(P \leq \alpha)$. Under the above reasoning, it follows from Equation (1) that if $\Pr(H_0 \mid P \leq \alpha)$ is high and yet p is low, then the Bayesian model fails the check at some small γ. That is argued with more rigor in Section 2.

If the priors behind the model are tentative, as is usually the case in Bayesian statistics, then it may be reasonable to insist that the Bayesian model used for inference does not fail the check. In that case, a low value of p would be inconsistent with a high value of $\Pr(H_0 \mid P \leq \alpha)$. That is made precise in the corollaries presented in Section 3. They support the intuition that null hypotheses with very low p values tend to have relatively low posterior probabilities unless the Bayesian model is inadequate as a predictor of the observation that $p \leq \alpha$. (By contrast, García-Donato and Chen (2005) and Gu, Hoijtink, and Mulder (2016) calibrated the Bayes factor to balance $\Pr(P \leq \alpha \mid H_0)$ with $\Pr(P \leq \alpha \mid H_1)$.)

Section 4 answers some reasonable objections:

1. Conditioning on the event of a low p-value was criticized by anonymous reviewers for its use of $\Pr(H_0 \mid P \leq \alpha)$ or $\Pr(H_0 \mid P \leq p)$ rather than $\Pr(H_0 \mid P = p)$ or $\Pr(H_0 \mid X = x)$ as the posterior probability. That criticism is answered in Section 4.1, which explains under what circumstances conditioning on $P \leq \alpha$ or $P \leq p$ may be warranted.
2. One of them also criticized the proposed approach's dependence on γ, the threshold determining at what point a Bayesian model is considered inadequate for its failure to predict the observed event that $P \leq \alpha$ or $P \leq p$. After explaining the origin of two default values of γ, Section 4.2 reviews arguments for and against using a threshold.

A discussion appears in Section 5, with the corollaries' implications for the practice of both Bayesian and frequentist hypothesis testing without setting a fixed significance level in Section 5.1 and with conclusions in Section 5.2.

2. Defense of Significance Testing on the Basis of Bayesian Model Checking

Some notation facilitates generalizing Examples 1 and 2 to other applications.

Let X denote a random sample from the probability density function f_θ for a θ in a set Θ of possible parameter values. A *Bayesian model* mdl is a pair $(\pi, \{f_\theta : \theta \in \Theta\})$, where π is a prior distribution of θ (Hill 1990; Bickel 2015). In short, $X \sim f_\vartheta$ and $\vartheta \sim \pi$ under model mdl, where ϑ is the unknown value of the parameter modeled as a random variable. According to model mdl, the observed sample x is a realization of X.

Pr will stand for the joint probability distribution of X and ϑ according to mdl. For example, if X and ϑ are scalars, then the joint probability that $X < 0$ and $\vartheta = 0$ is $\Pr(X < 0, \vartheta = 0) = \Pr(\{(y, 0) : y < 0\})$, and the marginal probability that $\vartheta = 0$ is $\Pr(\vartheta = 0) = \Pr(\{(y, 0) : y \in \mathbb{R}\})$, which was denoted by $\Pr(H_0)$ in Example 1 since its null hypothesis is $H_0 : \vartheta = 0$. More generally, for a $\theta_0 \in \Theta$, H_0 stands for the null hypothesis that $\vartheta = \theta_0$, and H_1 for the alternative hypothesis that $\vartheta \neq \theta_0$. $\Pr(H_0)$ and $\Pr(H_1)$ abbreviate $\Pr(\vartheta = \theta_0)$ and $\Pr(\vartheta \neq \theta_0)$, respectively.

Recall that a p-value tests a null hypothesis and has a uniform distribution between 0 and 1 if the null hypothesis is true. Analogously, a prior predictive p-value checks a Bayesian model and has a uniform distribution between 0 and 1 if the Bayesian model is true. Just as a sufficiently low p-value can lead to the rejection of a null hypothesis, a sufficiently low prior predictive p-value can lead to reformulating a Bayesian model. That analogy is exploited in the next two paragraphs.

The random variable P has these properties:

1. P, conditional on $\vartheta = \theta_0$, is uniformly distributed between 0 and 1, that is, $P \sim \mathrm{U}(0, 1)$ given that $X \sim f_{\theta_0}$. While the following results are stated for simplicity as if $P \sim \mathrm{U}(0, 1)$ were exact, they hold approximately for approximate p-values, including those for which P, conditional on $\vartheta = \theta_0$, converges to $\mathrm{U}(0, 1)$ in distribution.
2. P, conditional on $\vartheta \neq \theta_0$, is strictly stochastically less than a $\mathrm{U}(0, 1)$ random variable.

It follows that p, the realization of P, is an observed p-value for testing H_0 versus H_1.

Similarly, a prior predictive p-value for checking the model mdl is approximately distributed as $\mathrm{U}(0, 1)$ under mdl, that is, given that $X \sim f_\vartheta$ and $\vartheta \sim \pi$. Such a quantity may be constructed by calibrating P in the same way that posterior predictive p-values are calibrated (see, e.g., Hjort, Dahl, and Steinbakk 2006; Zhao and Xu 2014). Toward that end, let F denote the cumulative distribution function (CDF) of P given that $X \sim f_\vartheta$ and $\vartheta \sim \pi$. Unless $\Pr(H_0) = 1$, F is not the CDF of $\mathrm{U}(0, 1)$. However, since $F(P) \sim \mathrm{U}(0, 1)$ under mdl, the random variable $F(P)$ is a prior predictive p-value for checking mdl. The corresponding observed prior predictive p-value for checking mdl is

$$F(p) = \Pr(P \leq p).$$

In analogy with testing a null hypothesis, P plays the role of a test statistic, p plays the role of the observed value of P, and mdl plays the role of the tested null hypothesis.

If the p-value for testing H_0 is sufficiently low and yet the posterior probability of H_0 is sufficiently high, then the Bayesian model mdl fails the model check based on the prior predictive p-value and passes the check otherwise, as noted in Section 1. Recall that Section 1 argued that if the p-value is low and yet the posterior probability is not low relative to the prior probability, then mdl would fail its check. That is now made precise:

Theorem 1. If $p \leq \alpha$ and

$$\Pr(H_0 \mid P \leq \alpha) \geq \frac{\Pr(H_0)}{\gamma}\alpha, \tag{4}$$

then $F(p) \leq \gamma$.

Proof. By Bayes' theorem and by the assumptions that $p \leq \alpha$ and Equation (4),

$$F(p) = \Pr(P < p) = \frac{\Pr(P \leq p \mid H_0)\Pr(H_0)}{\Pr(H_0 \mid P \leq \alpha)} = \frac{p\Pr(H_0)}{\Pr(H_0 \mid P \leq \alpha)} \leq \frac{\alpha\Pr(H_0)}{\Pr(H_0 \mid P \leq \alpha)} \leq \frac{\alpha\Pr(H_0)}{\Pr(H_0)\alpha/\gamma} = \gamma.$$

□

The posterior probability of the null hypothesis does not necessarily have to be very high to satisfy Equation (4) and thereby trigger a failure of the model check:

Example 3. Returning to Example 1's $p = \Pr(|Z| \geq |z| \mid H_0) \approx 5 \times 10^{-7}$, make the conservative choices $\Pr(H_0) = 10/11$ and $\alpha = 5 \times 10^{-6}$. By Theorem 1, any Bayesian model mdl for which

$$\Pr(H_0 \mid P \leq \alpha) \gtrapprox \frac{10/11}{5 \times 10^{-3}} 5 \times 10^{-6} \approx 10^{-3}$$

would have a prior predictive p-value less than $\gamma = 5 \times 10^{-3}$.

3. Calibration of Significance Testing on the Basis of Bayesian Model Checking

While the upper bounds of this section are stated in terms of a significance level α to define what Section 1 means by a sufficiently low p-value, α can be optimized for the data rather than held fixed, as will be seen in Section 5.1.

The first corollary of Theorem 1 formally states Section 1's argument that if the p-value is sufficiently low, then any Bayesian model that passes the check yields a relatively low posterior probability of the null hypothesis.

Corollary 1. If $p \leq \alpha$ and $F(p) > \gamma$, then

$$\Pr(H_0 \mid P \leq \alpha) < \frac{\Pr(H_0)}{\gamma}\alpha. \tag{5}$$

Proof. Assume Equation (5) is false. Since that is equivalent to assuming Equation (4) is true, that assumption with $p \leq \alpha$ implies that $F(p) \leq \gamma$ (Theorem 1). That, however, contradicts the supposition that $F(p) > \gamma$. Therefore, Equation (5) cannot be false. □

Example 4. Plugging the numbers from Example 3 into Equation (5) reveals that any Bayesian model that passes the check at level 5×10^{-3} would generate a posterior probability of the null hypothesis no greater than about 10^{-3}.

The upper bound on $\Pr(H_0 \mid P \leq \alpha)$ given by Equation (5) depends on $\Pr(H_0)$. Two upper bounds that do not depend on $\Pr(H_0)$ are also available: one on the relative belief ratio, and the other on the Bayes factor. Considering the observation that $p \leq \alpha$ as evidence, the *relative belief ratio* (Evans 2015) favoring H_0 over H_1 under model mdl is

$$R(H_0 \mid P \leq \alpha) = \frac{\Pr(H_0 \mid P \leq \alpha)}{\Pr(H_0)}.$$

Whereas the posterior probability quantifies the degree to which the dataset as evidence for H_0 is sufficient for a conclusion

about H_0, the relative belief ratio quantifies the relevancy of the evidence to whether or not H_0 holds (Bickel 2018). The next result is an immediate consequence of Corollary 1.

Corollary 2. If $p \leq \alpha$ and $F(p) > \gamma$, then

$$R(H_0 \mid P \leq \alpha) < \frac{\alpha}{\gamma}.$$

Similarly, if $p \leq \alpha$, the *Bayes factor* favoring H_0 over H_1 is

$$B(P \leq \alpha) = \frac{\Pr(P \leq \alpha \mid H_0)}{\Pr(P \leq \alpha \mid H_1)}.$$

Like the relative belief ratio, the Bayes factor quantifies the relevancy rather than the sufficiency of the evidence (Lavine and Schervish 1999). It has the same prior-free upper bound.

Corollary 3. If $p \leq \alpha$ and $F(p) > \gamma$, then

$$B(P \leq \alpha) < \frac{\alpha}{\gamma}.$$

Proof. Since the Bayes factor is the ratio of the posterior odds to the prior odds,

$$\begin{aligned} B(P \leq \alpha) &= \frac{\Pr(H_0 \mid P \leq \alpha) / (1 - \Pr(H_0 \mid P \leq \alpha))}{\Pr(H_0) / (1 - \Pr(H_0))} \\ &= \frac{1 - \Pr(H_0)}{1 - \Pr(H_0) R(H_0 \mid P \leq \alpha)} R(H_0 \mid P \leq \alpha) \\ &\leq R(H_0 \mid P \leq \alpha), \end{aligned}$$

which, according to Corollary 2, is less than α/γ. □

Example 5. According to Corollaries 2 and 3, the settings $\alpha = 5 \times 10^{-6}$ and $\gamma = 5 \times 10^{-3}$ of Example 3 result in $\alpha/\gamma = 10^{-3}$ as the upper bound on both the relative belief ratio and the Bayes factor. That upper bound applies whether or not the prior probability of the null hypothesis is the $\Pr(H_0) = 10/11$ given in Example 3.

To use the corollaries to calibrate the p-value, their condition that $F(p) > \gamma$ needs to be consistent with p and $\Pr(H_0)$.

Theorem 2. The condition that $F(p) > \gamma$ is equivalent to each of these constraints:

$$\gamma < 1 - (1 - p) \Pr(H_0), \tag{6}$$

$$\Pr(H_0) < \frac{1 - \gamma}{1 - p}. \tag{7}$$

Proof. Let $F(\bullet \mid H_0)$ and $F(\bullet \mid H_1)$ denote the CDFs of P conditional on $\vartheta = \theta_0$ and $\vartheta \neq \theta_0$, respectively. $F(p) > \gamma$ if and only if

$$\begin{aligned} \gamma < F(p) &= \Pr(H_0) F(p \mid H_0) + \Pr(H_1) F(p \mid H_1) \\ &= \Pr(H_0) p + (1 - \Pr(H_0)) F(p \mid H_1) \\ &\leq \Pr(H_0) p + (1 - \Pr(H_0)), \end{aligned} \tag{8}$$

which is equivalent to constraint (6) and thus also to constraint (7). □

Each constraint has a different purpose. Constraint (6) prevents setting γ so high that no Bayesian model can pass the check without lowering its $\Pr(H_0)$. On the other hand, constraint (7) prevents using an excessively high $\Pr(H_0)$ with a corollary's upper bound based on a given value of γ.

Example 6. In the case of Example 3, the constraints of Theorem 2 take the form of

$$\begin{aligned} 5 \times 10^{-3} &< 1 - (1 - 5 \times 10^{-7}) \frac{10}{11} \approx \frac{1}{11} \\ \frac{10}{11} &< \frac{1 - 5 \times 10^{-3}}{1 - 5 \times 10^{-7}} \approx 1 - 5 \times 10^{-3}, \end{aligned}$$

each of which is obviously satisfied.

4. Can This Approach Survive the War on *p*-Values?

The approach of Sections 2 and 3 is open to attack in the spirit of recent criticisms of null hypothesis significance testing. Whereas Section 4.1 defends conditioning on a low value of p, Section 4.2 defends applying a threshold to $F(p) = \Pr(P \leq p)$, the prior predictive p-value.

4.1. When to Condition on a Low p-Value

When it is reasonable to use $\Pr(H_0 \mid P \leq \alpha)$, the posterior probability of the null hypothesis conditional on $P \leq \alpha$, rather than $\Pr(H_0 \mid X = x)$ or $\Pr(H_0 \mid P = p)$, its posterior probability conditional on $X = x$ or on $P = p$?

Example 7. Example 1, continued. Does a fully Bayesian analysis of the Higgs boson data require the detailed level of modeling that O'Hagan (2012) said would be needed to beat the 5-sigma test, complete with a range of choices of "prior distributions on the parameters of the background and the signal?" If the range of reasonable choices is too wide, the results would be inconclusive. To compute a range of plausible posterior probabilities would require transforming the raw dataset w generated by the experiment to some conceptually and computationally manageable sample x.

The priors could be further simplified by transforming the observation that $X = x$ to the observation that $P = p$. That data reduction would enable the use of any of several methods of transforming a p-value to a lower bound on $\Pr(H_0 \mid P = p)$ if the assumptions behind those methods are not too restrictive for the application at hand. (Held and Ott (2018) review many such methods.) However, even if the resulting *lower* bound ε is very small, that in itself would only warrant the conclusion that $\Pr(H_0 \mid P = p)$ is somewhere between ε and 1 (Sellke, Bayarri, and Berger 2001; Bickel 2019c; cf. Goodman 1999). That is a fundamental limitation of transforming a p-value to a range of values of $\Pr(H_0 \mid P = p)$, for a useful *upper* bound on $\Pr(H_0 \mid P = p)$ "usually does not exist" without stronger assumptions about the priors (Held and Ott 2018, p. 397).

By contrast, a useful upper bound on $\Pr(H_0 \mid P \leq \alpha)$ does exist, assuming only that the Bayesian model passes the check of Corollary 1. That upper bound on $\Pr(H_0 \mid P \leq \alpha)$ is what is

needed to claim that

$$\Pr(H_1 \mid P \le \alpha) = 1 - \Pr(H_0 \mid P \le \alpha) \ge 1 - \frac{\Pr(H_0)}{\gamma}\alpha,$$

by Equation (5). That arguably justifies reducing $X = x$ or $P = p$ to $P \le \alpha$ in the absence of reliable priors that would otherwise have to be specified to report an upper bound on $\Pr(H_0 \mid X = x)$ or $\Pr(H_0 \mid P = p)$.

An advantage of setting $p = \alpha$ and using $\Pr(H_0 \mid P \le p)$ as $\Pr(H_0 \mid P \le \alpha)$ is explained in Section 5.1. On the other hand, Colquhoun (2017, 2019) demonstrates that $\Pr(H_0 \mid P = p)$ is superior to $\Pr(H_0 \mid P \le p)$ given enough information about the Bayesian model. While such information is often lacking when only one null hypothesis is tested, it is present in the set of p-values when many null hypotheses are tested. In that case, estimates of $\Pr(H_0 \mid P \le p)$ perform worse than those of $\Pr(H_0 \mid P = p)$ when targeting $\chi(H_0)$, which is 1 when H_0 is true and 0 when H_0 is false (Bickel and Rahal 2019). In the case of multiple testing, estimates of $\Pr(H_0 \mid P \le p)$ may be transformed to estimates of $\Pr(H_0 \mid P = p)$ (Bickel 2019a, chap. 6; Bickel and Rahal 2019).

For an example of a single-test transformation of a p-value to a lower bound on $\Pr(H_0 \mid P = p)$ that generalizes to estimates of $\Pr(H_0 \mid P = p)$ for multiple p-values, see Bickel (2019a, chap. 7). For another book on multiple testing that emphasizes $\Pr(H_0 \mid P = p)$, see Efron (2010).

4.2. The Model-Checking Threshold γ

Because 0.05 is the conventional p-value threshold α for hypothesis testing, it is also the most common prior predictive p-value threshold γ for checking Bayesian models. Since Benjamin et al. (2017), following Johnson (2013), suggested lowering α to 0.005 partly on the basis of transformations of p-values to lower bounds on Bayes factors, $\gamma = 5 \times 10^{-3}$ is used in some examples in this article. However, the other argument that Benjamin et al. (2017) made for the $\alpha = 5 \times 10^{-3}$ threshold is based on meta-analysis of studies in certain fields that rely heavily on significance testing, an argument that is irrelevant to choosing a value of the threshold γ for Bayesian model checking. The reasoning of this article based on $\gamma = 5 \times 10^{-3}$ would only be strengthened by increasing the threshold to $\gamma = 5 \times 10^{-2}$, for that change would decrease the upper bound on $\Pr(H_0 \mid P \le p)$ by a factor of 10, as seen in Equation (9) and Figure 1.

The need to choose a threshold γ for Bayesian model checking seems to be a weakness of the procedure of checking a model and revising it if the model fails the check, for the value of the threshold is largely arbitrary. Another apparent disadvantage of the procedure is its discontinuity: a model suddenly jumps from being considered completely adequate to being considered completely inadequate as soon as γ is transgressed.

On the other hand, just as Wellek (2017, sec. 3.4) and Mayo (2019) have argued for the need for thresholds in hypotheses testing, they may serve similar ends in checking Bayesian models. The purpose of the latter is to determine whether a model requires revision (e.g., Evans 2015). The dichotomous decision of either revising or not revising a model lends itself to a threshold. That said, there is no reason to insist on a particular fixed threshold for all applications.

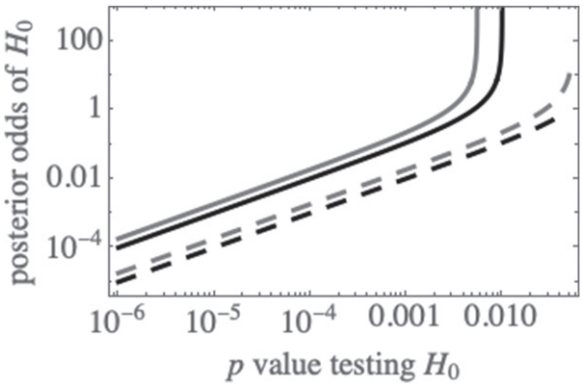


Figure 1. Upper bounds of the posterior odds that the data came from H_0, as a function of p. $\Pr(H_0)$ is 1/2 (black) or 10/11 (gray, following Benjamin et al. (2017)); γ is 5×10^{-3} (solid) or 5×10^{-2} (dashed).

5. Discussion

Section 5.1 features $(\Pr(H_0)/\gamma)\,p$ as the upper bound on the posterior probability of the null hypothesis. Section 5.2 then offers some simple conclusions.

5.1. Implications for Hypothesis Testing in Bayesian and Frequentist Practice

Whereas Theorem 1 has direct implications for Bayesian hypothesis testing, its corollaries and Theorem 2 have implications for frequentist hypothesis testing.

If a Bayesian data analysis does not yield a low posterior probability of the null hypothesis H_0, the result may be checked first by computing p, a p-value testing H_0. If it is low, then the next step is to formally check the Bayesian model using a prior predictive p-value. Theorem 1 provides an unusually simple way to perform that check. The main limitation from a Bayesian perspective might appear to be the fact that the posterior in Theorem 1 conditions on $P \le \alpha$ rather than on $X = x$, but see Section 4.1.

Frequentist hypothesis testing yields a p-value in need of careful interpretation (Wasserstein and Lazar 2016). The corollaries of the theorem provide simple ways to transform the p-value to upper bounds on the posterior probability of H_0 and on related, prior-free quantities under an unspecified Bayesian model mdl. The upper bounds depend on the significance level α, but fixing its value raises concerns (e.g., Naaman 2016; Wasserstein and Lazar 2016). Fortunately, the value of α may instead be optimized in light of the data. Specifically, taking the least upper bound of each corollary's upper bound results in p as the optimal value of α. Corollaries 1, 2, and 3 then suggest using $(\Pr(H_0)/\gamma)\,p$, $(1/\gamma)\,p$, and $(1/\gamma)\,p$ as the optimal upper bounds for the posterior probability, relative belief ratio, and Bayes factor, respectively:

$$\Pr(H_0 \mid P \le p) < \frac{\Pr(H_0)}{\gamma}\,p = \Pr(H_0 \mid P \le p) = \inf_{\alpha \ge p} \frac{\Pr(H_0)}{\gamma}\alpha$$
$$R(H_0 \mid P \le p) < \frac{1}{\gamma}\,p = R(H_0 \mid P \le p) = \inf_{\alpha \ge p} \frac{\alpha}{\gamma} \tag{9}$$

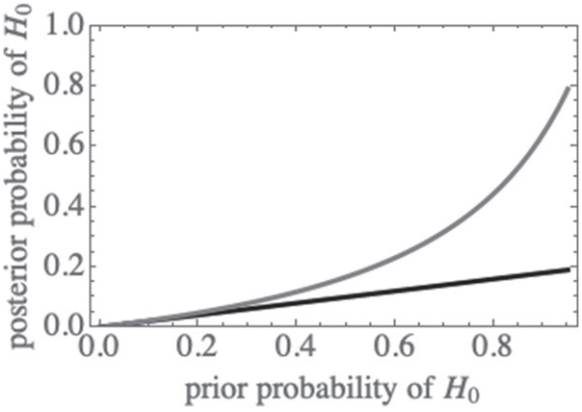


Figure 2. Upper bounds of the posterior probability of H_0 as a function of the prior probability $\Pr(H_0)$ given $p = 10^{-3}$ and $\gamma = 5 \times 10^{-3}$. The gray curve is derived from $B(P \leq p)$ and Bayes' theorem; the black curve is $\Pr(H_0 \mid P \leq p) = \Pr(H_0) R(H_0 \mid P \leq p)$.

$$B\left(P \leq p\right) < \frac{1}{\gamma} p = B\left(P \leq p\right) = \inf_{\alpha \geq p} \frac{\alpha}{\gamma},$$

where γ satisfies Theorem 2's constraint (6).

For the first upper bound, Figure 1 displays $\Pr\left(H_0 \mid P \leq p\right) / \left(1 - \Pr\left(H_0 \mid P \leq p\right)\right)$ for different values of p, $\Pr(H_0)$, and γ. While the results certainly depend on the choices of $\Pr(H_0)$ and γ, the overall message is clear: low p-values tend to lead to low posterior probabilities that the data came from a distribution consistent with the null hypothesis.

Not requiring the prior probability $\Pr(H_0)$, either of the other two upper bounds, $R\left(H_0 \mid P \leq p\right)$ and $B\left(P \leq p\right)$, might be reported to allow each reader to combine it with a different prior if necessary. Objectivity in that sense is often cited as a reason to report Bayes factors rather than posterior probabilities (e.g., Wellcome Trust Case Control Consortium 2007; Bickel 2019b). However, Figure 2 suggests that $B\left(P \leq p\right)$, the Bayes factor bound, is too conservative unless $\Pr(H_0) \leq 1/2$.

For that reason, $R\left(H_0 \mid P \leq p\right)$, the optimal upper bound on the relative belief ratio, is preferable as a prior-free summary of the test result. Each recipient of the report may easily multiply $R\left(H_0 \mid P \leq p\right)$ by a hypothetical or estimated value of $\Pr(H_0)$ that satisfies constraint (7). The resulting product is $\Pr\left(H_0 \mid P \leq p\right)$, the optimal upper bound on the posterior probability.

If $\Pr\left(H_0 \mid P \leq p\right)$ is sufficiently low, the null hypothesis may be considered false for decision making purposes since its posterior probability according to a model mdl passing the check is even lower. When warranted, that can be formalized in terms of minimizing posterior expected loss or, considering $\left[0, \Pr\left(H_0 \mid P \leq p\right)\right[$ as an imprecise probability, in terms of a generalization of minimizing posterior expected loss (e.g., Troffaes 2007).

Example 8. Example 3, continued. Let mdl be any Bayesian model that passes the model check. Since constraint (7) is satisfied, Corollary 1 yields Equation (9) and thus

$$\Pr\left(H_0 \mid P \lessapprox 5 \times 10^{-7}\right) \lessapprox \frac{10/11}{5 \times 10^{-3}} 5 \times 10^{-7} \approx 10^{-4},$$

indicating that the low p-value is not illusory.

Example 9. Let X denote the number of successes out of n independent Bernoulli trials, each with an unknown probability θ of success. Thus, the law of X is the binomial distribution with parameters n and θ. Bernardo (2011) considered the null hypothesis $H_0 : \theta = 1/2$ with $x = 52{,}263{,}471$ observed successes among $n = 104{,}490{,}000$ trials. Bernardo (2011) interpreted the null hypothesis as the claim that there is no measured effect due to extrasensory perception (ESP) and that there is not even a very small systematic error in the measurements. In the discussion, Luis Pericchi reported that, in spite of the p-value of 3×10^{-4}, the posterior probability $\Pr(H_0 \mid X = x)$ would be over 95% if $\Pr(H_0) = 1/2$ (Bernardo 2011). By contrast, following the procedure of Example 8 with $\gamma = 5 \times 10^{-3}$, formula (9) yields

$$\begin{aligned} \Pr\left(H_0 \mid P \leq 3 \times 10^{-4}\right) &\lessapprox \frac{5 \times 10^{-1}}{5 \times 10^{-3}} 3 \times 10^{-4} \\ &\approx 3 \times 10^{-2} = 0.03, \end{aligned} \tag{10}$$

indicating a high posterior probability that there is some systematic error for any model passing the check at level γ. Thus, the model yielding $\Pr(H_0 \mid X = x) > 95\%$ would miserably fail the check since $0.95 \gg 0.03$. The case of Equation (10) obeys constraint (7):

$$\Pr(H_0) < \frac{1 - \gamma}{1 - p} = \frac{1 - 5 \times 10^{-3}}{1 - 3 \times 10^{-4}} = 99.5\%. \tag{11}$$

If the systematic bias could be ruled out, then H_0 would say there is no measured effect due to ESP, in which case $\Pr(H_0)$ would be closer to 100% than is allowed by Equation (11), and thus formula (9) would not apply.

5.2. Conclusions

This article both defends and calibrates null hypothesis significance testing. The defense counters the attack made on the basis that a very low p-value does not necessarily imply the posterior probability of the null hypothesis is low relative to its prior probability. While there certainly are Bayesian models for which that is true, those models may be found inadequate by a prior predictive check since they fail to predict that the p-value would be as low as it is (Theorem 1).

It does not follow that a p-value slightly under 0.05 warrants concluding that the posterior probability of the null hypothesis is small. Rather, the same prior predictive p-value that defends null hypothesis significance testing also leads to two calibrations of the p-value that tests the null hypothesis.

The main calibration is $(\Pr(H_0)/\gamma)\, p$, an upper bound on $\Pr(H_0 \mid P \leq \alpha)$ as the posterior probability of the null hypothesis, according to Equation (9).

Acknowledgments

Remarks by two anonymous referees resulted in the addition of Sections 4 and 5.2 and in other improvements. I also thank an associate editor and Daniel Jeske for guidance leading to increased elegance, concision, and clarity.

Funding

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009).

References

Bartlett, M. S. (1957), "A Comment on D. V. Lindley's Statistical Paradox," *Biometrika*, 44, 533–534. [249]

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2017), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6. [250,253]

Bernardo, J. M. (2011), "Integrated Objective Bayesian Estimation and Hypothesis Testing," *Bayesian Statistics*, 9, 1–68. [254]

Bickel, D. R. (2015), "Inference After Checking Multiple Bayesian Models for Data Conflict and Applications to Mitigating the Influence of Rejected Priors," *International Journal of Approximate Reasoning*, 66, 53–72. [250]

——— (2018), "Confidence Distributions and Empirical Bayes Posterior Distributions Unified as Distributions of Evidential Support," working paper, DOI: 10.5281/zenodo.2529438. [252]

——— (2019a), *Genomics Data Analysis: False Discovery Rates and Empirical Bayes Methods*, New York: Chapman and Hall/CRC, available at *https://davidbickel.com/genomics/*. [253]

——— (2019b), "Reporting Bayes Factors or Probabilities to Decision Makers of Unknown Loss Functions," *Communications in Statistics—Theory and Methods*, 48, 2163–2174. [254]

——— (2019c), "Sharpen Statistical Significance: Evidence Thresholds and Bayes Factors Sharpened Into Occam's Razor," *Stat*, 8, e215. [252]

Bickel, D. R., and Rahal, A. (2019), "Correcting False Discovery Rates for Their Bias Toward False Positives," *Communications in Statistics—Simulation and Computation*, DOI: 10.1080/03610918.2019.1630432. [253]

Colquhoun, D. (2017), "The Reproducibility of Research and the Misinterpretation of *p*-Values," *Royal Society Open Science*, 4, 171085. [253]

——— (2019), "The False Positive Risk: A Proposal Concerning What to Do About *p*-Values," *The American Statistician*, 73, 192–201. [253]

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge: Cambridge University Press. [253]

Evans, M. (2015), *Measuring Statistical Evidence Using Relative Belief*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, New York: CRC Press. [251,253]

Fisher, R. A. (1973), *Statistical Methods and Scientific Inference*, New York: Hafner Press. [249]

García-Donato, G., and Chen, M.-H. (2005), "Calibrating Bayes Factor Under Prior Predictive Distributions," *Statistica Sinica*, 15, 359–380. [250]

Goodman, S. N. (1999), "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013. [252]

Gu, X., Hoijtink, H., Mulder, J. (2016), "Error Probabilities in Default Bayesian Hypothesis Testing," *Journal of Mathematical Psychology*, 72, 130–143. [250]

Held, L., and Ott, M. (2018), "On *p*-Values and Bayes Factors," *Annual Review of Statistics and Its Application*, 5, 393–419. [252]

Hill, J. R. (1990), "A General Framework for Model-Based Statistics," *Biometrika*, 77, 115–126. [250]

Hjort, N., Dahl, F., and Steinbakk, G. (2006), "Post-Processing Posterior Predictive *p* Values," *Journal of the American Statistical Association*, 101, 1157–1174. [251]

Johnson, V. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences of the United States of America*, 110, 19313–19317. [253]

Lavine, M., and Schervish, M. J. (1999), "Bayes Factors: What They Are and What They Are Not," *American Statistician*, 53, 119–122. [252]

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192. [249]

Mayo, D. G. (2019), "*P*-value Thresholds: Forfeit at Your Peril," *European Journal of Clinical Investigation*, 49, e13170. [253]

Micheas, A. C., and Dey, D. K. (2003), "Prior and Posterior Predictive *p*-Values in the One-Sided Location Parameter Testing Problem," *Sankhya*, 65, 158–178. [250]

Naaman, M. (2016), "Almost Sure Hypothesis Testing and a Resolution of the Jeffreys-Lindley Paradox," *Electronic Journal of Statistics*, 10, 1526–1550. [253]

O'Hagan, A. (2012), "Higgs Boson—Digest and Discussion," unpublished document, *http://tonyohagan.co.uk/academic/pdf/HiggsBoson.pdf*. [249,252]

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of *p* Values for Testing Precise Null Hypotheses," *American Statistician*, 55, 62–71. [252]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [249]

Troffaes, M. C. M. (2007), "Decision Making Under Uncertainty Using Imprecise Probabilities," *International Journal of Approximate Reasoning*, 45, 17–29. [254]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [249,253]

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, 661–678. [254]

Wellek, S. (2017), "A Critical Evaluation of the Current '*p*-Value Controversy'," *Biometrical Journal*, 59, 854–872. [253]

Zhao, G., and Xu, X. (2014), "The One-Sided Posterior Predictive *p*-Value for Fieller's Problem," *Statistics and Probability Letters*, 95, 57–62. [251]