# Null hypothesis significance testing defended and calibrated by Bayesian model checking

David R. Bickel

# Null hypothesis significance testing defended and calibrated by Bayesian model checking

David R. Bickel

December 31, 2018

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670

dbickel@uottawa.ca

**Abstract**

Null hypothesis significance testing is often criticized because attaining statistical significance does not necessarily imply that the posterior probability is low relative to to the prior probability. That discrepancy between significance testing and Bayesian hypothesis testing occurs whenever the statistical power is low.

However, if both the significance level and the power are small, then the Bayesian model would assign a low prior probability to the observation that the data achieved statistical significance. That conflict between the model and the data may indicate that the model needs revision. More formally, if the achieved significance level is sufficiently small while the posterior probability is insufficiently small, then the model will fail a prior predictive model check, being found inadequate for inference and decision purposes.

That result leads to a simple way to calibrate a $p$ value by transforming it into an upper bound on the posterior probability of the null hypothesis for any Bayesian model that would pass the prior predictive check. The calibration may be calculated from a prior probability of the null hypothesis and the stringency of the prior predictive check without more detailed Bayesian modeling. An advantage of an upper bound as opposed to the usual lower bounds is that it justifies concluding that the null hypothesis has a low posterior probability.

**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.** Regarding a 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?"

Hypothetically, the discovery could well be illusory with very high probability, for a traditional Bayesian analysis with strictly positive probability of a simple null hypothesis (one with all of its mass at a point in hypothesis space) and a diffuse alternative hypothesis can yield a very high posterior probability of the null hypothesis in spite of the low $p$ value (Lindley, 1957). ▲

That phenomenon of a high posterior probability of the null hypothesis relative to its prior probability in spite of a very low $p$ value is called the Jeffreys-Lindley paradox; see Cousins (2017) for a review. The root of the paradox is exposed by a normal model that approximates many more complex Bayesian models and that Held and Ott (2016) and others have used to calibrate $p$ values.

**Example 2.** Consider a normal random variable $X$ of unit mean and unknown standard deviation $\theta$. The null hypothesis is $H_0 : \theta = 1$, the alternative hypothesis is $H_1 : \theta = 1.1$, and the observed value of $X$ is $x = 3$, which implies that the two-sided $p$ value is 0.003, just under $\alpha = 0.005$, the significance level recommended by Benjamin et al. (2017). However, even when reducing the observation that $X = x$ to the observation that $p\left(H_0; X\right) \leq \alpha$, the posterior probability of $H_0$ is not necessarily very low compared to its prior probability.

Rather, Bayes's theorem says the posterior odds of $H_0$ is

$$\frac{\text{Prob}\left(H_0 | \, p\left(H_0; X\right) \leq \alpha\right)}{\text{Prob}\left(H_1 | \, p\left(H_1; X\right) \leq \alpha\right)} = \frac{\text{Prob}\left(H_0\right)}{\text{Prob}\left(H_1\right)} \frac{\text{Prob}\left(p\left(H_0; X\right) \leq \alpha | H_0\right)}{\text{Prob}\left(p\left(H_0; X\right) \leq \alpha | H_1\right)} = \frac{\text{Prob}\left(H_0\right)}{\text{Prob}\left(H_1\right)} \frac{\alpha}{\text{power}\left(\alpha\right)}, \tag{1}$$

which is the product of prior odds of $H_0$ and the Bayes factor, where $\text{power}\left(\alpha\right)$ is the statistical power, the probability of the observation that $p\left(H_0; X\right) \leq \alpha$ conditional on $H_1$. In this case, $\text{power}\left(\alpha\right) = 0.11$. If the prior odds is 10, an empirically supported default (Benjamin et al., 2017), then the posterior odds is about $10\left(0.005/0.1\right) = 1:2$ by equation (1), and thus $\text{Prob}\left(H_0 | \, p\left(H_0; X\right) \leq \alpha\right) \approx 1/3$. ▲

Applying generally, equation (1) reveals two facts relevant to the relationship between posterior probability and the $p$ value. First, the Bayes factor measures the posterior odds relative to the prior odds and thus, indirectly, the posterior probability relative to the prior probability. Second, the Bayes factor increases with $\alpha$ and decreases with $\text{power}\left(\alpha\right)$. Putting both facts together, observing a $p$ value less than a very low significance level is compatible with a relatively high posterior probability if the statistical power is low. That effect of power is illustrated graphically in Trafimow (2003), which records much of the reasoning behind the $p$-value ban (Trafimow and Marks, 2015).

On the other hand, when both the significance level and the power are low, $\text{Prob}\left(p\left(H_0; X\right) \leq \alpha\right)$ is also low, for

$$\text{Prob}\left(p\left(H_0; X\right) \leq \alpha\right) = \text{Prob}\left(H_0\right)\alpha + \text{Prob}\left(H_1\right)\text{power}\left(\alpha\right),$$

which means the Bayesian model behind $\text{power}\left(\alpha\right)$ did a poor job predicting the observation that $p\left(H_0; X\right) \leq \alpha$. Thus, when a null hypothesis achieves a very low $p$ value and yet not a relatively low posterior probability, there may be substantial disagreement between the model and the data, suggesting that the model be revised in the direction of the data. The underlying principle is to use models suitable for inference and decision making that are reasonably consistent with the data. That principle is often formalized in terms of using prior predictive $p$ values to check Bayesian models for how well they predict the data (e.g., Micheas and Dey, 2003).

A simple way to perform such a prior predictive model check requires little more information than a $p$ value that tests $H_0$, a prior probability of $H_0$, and the stringency of the check (Sec. 2). That check supports 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 $p\left(H_0; X\right) \leq \alpha$ (Theorem 1).

Some corollaries for calibrating the $p$ value appear in Section 3.

Implications for the practice of both Bayesian and frequentist hypothesis testing without setting a fixed significance level are discussed in Section 4. Its $p$ value calibration provides an *upper* bound of the posterior probability of $H_0$ instead of the *lower* bound provided by the methods reviewed in Held and Ott (2018). A small upper bound has the advantage that it enables the conclusion that $H_0$ is improbable in light of the data, whereas a small lower bound would only warrant the conclusion that $H_0$ *could be* improbable in light of the data (Sellke et al., 2001; Bickel, 2018c).

## 2 Defense of significance testing on the basis of Bayesian model checking

Let $X$ denote a random sample from the probability density function $f_\theta$ for a $\theta$ in a set $\Theta$ of possible parameter values. A *Bayesian model* $M$ is a pair $(\pi, \{f_\theta : \theta \in \Theta\})$, where $\pi$ is a prior distribution of $\theta$ (Hill, 1990; Bickel, 2015). In short, $X \sim f_\vartheta$ and $\vartheta \sim \pi$ under model $M$, where $\vartheta$ is the true value of the parameter as a random variable. $M$ says the observed sample $x$ is a realization of $X$. $P_M$ will stand for the joint probability distribution of $X$ and $\vartheta$ according to $M$.

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 $\vartheta \neq \theta_0$. $P_M(H_0)$ and $P_M(H_1)$ abbreviate $P_M(\vartheta = \theta_0)$ and $P_M(\vartheta \neq \theta_0)$, respectively.

The function $p(H_0; \bullet)$ has these properties:

1. The random variable $p(H_0; X)$, conditional on $\vartheta = \theta_0$, is uniformly distributed between 0 and 1, that is, $p(H_0; X) \sim \mathrm{U}(0, 1)$ given that $X \sim f_{\theta_0}$.

2. The random variable $p(H_0; X)$, conditional on $\vartheta \neq \theta_0$, is strictly stochastically less than a $\mathrm{U}(0, 1)$ random variable.

It follows that $p(H_0; x)$ is an observed $p$ value for testing $H_0$ versus $H_1$. While the following results are stated for simplicity as if $p(H_0; X) \sim \mathrm{U}(0, 1)$ were exact, they hold approximately for approximate $p$ values, including those for which $p(H_0; X)$, conditional on $\vartheta = \theta_0$, converges to $\mathrm{U}(0, 1)$ in distribution.

Similarly, a prior predictive $p$ value for checking the model $M$ is approximately distributed as $\mathrm{U}(0, 1)$ under $M$, that is, given that $X \sim f_\vartheta$ and $\vartheta \sim \pi$. Such a quantity may be constructed by calibrating $p(H_0; X)$

in the same way that posterior predictive $p$ values are calibrated; see, for example, Hjort et al. (e.g., 2006) and Zhao and Xu (e.g., 2014). Toward that end, let $F_M$ denote the cumulative distribution function (CDF) of $p(H_0; X)$ given that $X \sim f_\vartheta$ and $\vartheta \sim \pi$. Unless $P_M(H_0) = 1$, $F_M$ is not the CDF of $\mathrm{U}(0,1)$. However, since $F_M(p(H_0; X)) \sim \mathrm{U}(0,1)$ under $M$, the random variable $F_M(p(H_0; X))$ is a prior predictive $p$ value for checking $M$ and is appropriately written as $p^{\mathrm{pred}}(M; X)$. The corresponding observed prior predictive $p$ value for checking $M$ is $p^{\mathrm{pred}}(M; x) = F_M(p(H_0; x))$.

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 fails the model check based on the prior predictive $p$ value.

**Theorem 1.** *If $p(H_0; x) \leq \alpha$, then $p^{\mathrm{pred}}(M; x) \leq \gamma$ for every $\gamma \in ]0,1]$ such that*

$$P_M(H_0 | \, p(H_0; X) \leq \alpha) \geq \frac{P_M(H_0)}{\gamma}\alpha. \tag{2}$$

*Proof.* For any $\gamma \in ]0,1]$ satisfying equation (2),

$$\begin{aligned}
\frac{P_M(H_0)}{\gamma}\alpha &\leq P_M(H_0 | \, p(H_0; X) \leq \alpha) \\
&= \frac{P_M(H_0) \, P_M(p(H_0; X) \leq \alpha | \vartheta = \theta_0)}{P_M(p(H_0; X) \leq \alpha)} \\
&= \frac{P_M(H_0) \, F_M(\alpha | H_0)}{P_M(H_0) \, F_M(\alpha | H_0) + P_M(H_1) \, F_M(\alpha | H_1)},
\end{aligned}$$

where $F_M(\bullet | H_0)$ and $F_M(\bullet | H_1)$ are the CDFs of $p(H_0; X)$ conditional on $\vartheta = \theta_0$ and $\vartheta \neq \theta_0$, respectively. (The asymmetry is intended, for no $p(H_1; X)$ is necessary.) Therefore, by the definition of $p^{\mathrm{pred}}(M; x)$, by $p(H_0; x) \leq \alpha$, and by $p(H_0; X) \sim \mathrm{U}(0,1)$ conditional on $\vartheta = \theta_0$,

$$\begin{aligned}
p^{\mathrm{pred}}(M; x) &= F_M(p(H_0; x)) \\
&= P_M(H_0) \, F_M(p(H_0; x) | H_0) + P_M(H_1) \, F_M(p(H_0; x) | H_1) \\
&\leq P_M(H_0) \, F_M(\alpha | H_0) + P_M(H_1) \, F_M(\alpha | H_1) \\
&\leq \frac{P_M(H_0) \, F_M(\alpha | H_0)}{P_M(H_0) \, \alpha / \gamma} = \frac{\alpha}{\alpha / \gamma} = \gamma.
\end{aligned} \tag{3}$$

□

**Example 3.** Returning to Example 1's $p\left(H_0;x\right)\approx 5\times 10^{-7}$, make the conservative choices $P_M\left(H_0\right)=10/11$ and $\alpha=5\times 10^{-6}$. By Theorem 1, any Bayesian model $M$ for which

$$P_M\left(H_0|\, p\left(H_0;X\right)\leq\alpha\right)\gtrapprox\frac{10/11}{5\times 10^{-3}}5\times 10^{-6}\approx 10^{-3}$$

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 calibrations of this section are stated in terms of a significance level $\alpha$, that level need not be fixed but can be optimized for the data, as will be seen in Section 4.

**Corollary 1.** *If $p\left(H_0;x\right)\leq\alpha$ for an $\alpha\in\left]0,1\right]$ and $p^{\text{pred}}\left(M;x\right)>\gamma$ for a $\gamma\in\left]0,1\right]$, then*

$$P_M\left(H_0|\, p\left(H_0;X\right)\leq\alpha\right)<\frac{P_M\left(H_0\right)}{\gamma}\alpha. \tag{4}$$

*Proof.* According to Theorem 1,

$$p\left(H_0;x\right)\leq\alpha \text{ and } \neg\left(P_M\left(H_0|\, p\left(H_0;X\right)\leq\alpha\right)<\frac{P_M\left(H_0\right)}{\gamma}\alpha\right)\implies\neg\left(p^{\text{pred}}\left(M;x\right)>\gamma\right). \tag{5}$$

Since $p\left(H_0;x\right)\leq\alpha$ and $p^{\text{pred}}\left(M;x\right)>\gamma$ by assumption, equation (5) can only be true if equation (4) is true. □

Two upper bounds that do not depend on $P_M\left(H_0\right)$ are also available: one on the relative belief ratio, and the other on the Bayes factor. If $p\left(H_0;x\right)\leq\alpha$ , the *relative belief ratio* (Evans, 2015) favoring $H_0$ over $H_1$ under model $M$ is

$$R_M\left(H_0|\, p\left(H_0;X\right)\leq\alpha\right)=\frac{P_M\left(H_0|\, p\left(H_0;X\right)\leq\alpha\right)}{P_M\left(H_0\right)}.$$

Whereas the posterior probability quantifies the degree to which the data set 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, 2018a). The next result is an immediate consequence of Corollary 1.

**Corollary 2.** *If* $p(H_0;x) \leq \alpha$ *for an* $\alpha \in ]0,1]$ *and* $p^{\text{pred}}(M;x) > \gamma$ *for a* $\gamma \in ]0,1]$*, then*

$$R_M(H_0 | p(H_0;X) \leq \alpha) < \frac{\alpha}{\gamma}.$$

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

$$B_M(p(H_0;X) \leq \alpha) = \frac{P_M(p(H_0;X) \leq \alpha | H_0)}{P_M(p(H_0;X) \leq \alpha | H_1)}.$$

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* $p(H_0;x) \leq \alpha$ *for an* $\alpha \in ]0,1]$ *and* $p^{\text{pred}}(M;x) > \gamma$ *for a* $\gamma \in ]0,1]$*, then*

$$B_M(p(H_0;X) \leq \alpha) < \frac{\alpha}{\gamma}.$$

*Proof.* By Corollary 1,

$$\frac{P_M(H_0)\alpha/\gamma}{1 - P_M(H_0)\alpha/\gamma} > \frac{P_M(H_0 | p(H_0;X) \leq \alpha)}{P_M(H_1 | p(H_0;X) \leq \alpha)} = \frac{P_M(H_0)}{P_M(H_1)} \frac{P_M(p(H_0;X) \leq \alpha | H_0)}{P_M(p(H_0;X) \leq \alpha | H_1)}$$

$$\begin{aligned}
\therefore \frac{P_M(p(H_0;X) \leq \alpha | H_0)}{P_M(p(H_0;X) \leq \alpha | H_1)} &< \frac{P_M(H_1)}{P_M(H_0)} \frac{P_M(H_0)\alpha/\gamma}{1 - P_M(H_0)\alpha/\gamma} = \frac{P_M(H_1)\alpha/\gamma}{1 - (1 - P_M(H_1))\alpha/\gamma} \\
&= \frac{P_M(H_1)\alpha/\gamma}{1 - \alpha/\gamma + P_M(H_1)\alpha/\gamma} = \frac{1}{1 + \frac{1-\alpha/\gamma}{P_M(H_1)\alpha/\gamma}} = \frac{1}{1 + \frac{\gamma-\alpha}{P_M(H_1)\alpha}} \\
&\leq \frac{1}{1 + \frac{\gamma-\alpha}{\alpha}} = \frac{1}{\gamma/\alpha} = \frac{\alpha}{\gamma}.
\end{aligned}$$

□

To use the corollaries to calibrate the $p$ value $p(H_0;x)$, their condition that $p^{\text{pred}}(M;x) > \gamma$ needs to be consistent with $p(H_0;x)$ and $P_M(H_0)$.

**Theorem 2.** *Consider a* $P_M(H_0) \in [0,1]$*, a* $p(H_0;x) \in ]0,1]$*, a* $\gamma \in ]0,1]$*, and a conditional probability distribution* $P_M(\bullet|H_0)$*. There is a conditional probability distribution* $P_M(\bullet|H_1)$ *that satisfies* $p^{\text{pred}}(M;x) > \gamma$ *if and only if it also satisfies these equivalent constraints:*

$$\gamma < 1 - (1 - p(H_0;x)) P_M(H_0) \tag{6}$$

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

*Proof.* By equation (3), an $P_M(\bullet|H_1)$ such that $p^{\text{pred}}(M;x) > \gamma$ exists if and only if

$$\begin{aligned}\gamma &< P_M(H_0) F_M(p(H_0;x)|H_0) + P_M(H_1) F_M(p(H_0;x)|H_1) \\ &= P_M(H_0) p(H_0;x) + (1 - P_M(H_0)) F_M(p(H_0;x)|H_1)\end{aligned}$$

and thus if and only if a $\beta \in [0,1]$ exists such that

$$\gamma < P_M(H_0) p(H_0;x) + (1 - P_M(H_0))(1-\beta).$$

That assertion of existence is equivalent to

$$\begin{aligned}\gamma &< \sup_{\beta \in [0,1]} P_M(H_0) p(H_0;x) + (1 - P_M(H_0))(1-\beta) \\ &= P_M(H_0) p(H_0;x) + 1 - P_M(H_0),\end{aligned}$$

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

Each constraint has a different purpose. Constraint (6) prevents setting $\gamma$ so high that no Bayesian model can pass the check without lowering its $P_M(H_0)$. On the other hand, constraint (7) prevents using a $P_M(H_0)$ that is too close to 1 with a corollary's upper bound based on a given value of $\gamma$.

# 4 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 $p\left(H_0; x\right)$, 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 is that the posterior in Theorem 1 conditions on $p\left(H_0; X\right) \leq \alpha$ rather than on $X = x$.

Frequentist hypothesis testing yields a $p$ value in need of careful interpretation (Wasserstein and Lazar, 2016). The corollaries of the theorem provide equally simple ways to transform the $p$ value to upper bounds on the posterior probability of $H_0$ and on related, prior-free quantities under a Bayesian model $M$ that need not be specified. The upper bounds depend on the significance level $\alpha$, which need not be fixed ahead of time contra recommendations of the American Statistical Association (Wasserstein and Lazar, 2016), but may instead be optimized in light of the data. Specifically, taking the least upper bound of each corollary's upper bound results in $\alpha^{\text{opt}} = p\left(H_0; x\right)$ as the optimal value of $\alpha$. Corollaries 1, 2, and 3 then suggest using $\left(P_M\left(H_0\right)/\gamma\right) p\left(H_0; x\right)$, $\left(1/\gamma\right) p\left(H_0; x\right)$, and $\left(1/\gamma\right) p\left(H_0; x\right)$ as the optimal upper bounds for the posterior probability, relative belief ratio, and Bayes factor, respectively:

$$\begin{aligned} P_M\left(H_0 \middle| \, p\left(H_0; X\right) \leq p\left(H_0; x\right)\right) &< \frac{P_M\left(H_0\right)}{\gamma} p\left(H_0; x\right) \\ &= P_M^{\text{upper}}\left(H_0 \middle| \, p\left(H_0; X\right) \leq \alpha^{\text{opt}}\right) \overset{\text{def}}{=} \inf_{\alpha \geq p(H_0; x)} \frac{P_M\left(H_0\right)}{\gamma} \alpha \end{aligned} \tag{8}$$

$$\begin{aligned} R_M\left(H_0 \middle| \, p\left(H_0; X\right) \leq p\left(H_0; x\right)\right) &< \frac{1}{\gamma} p\left(H_0; x\right) \\ &= R_M^{\text{upper}}\left(H_0 \middle| \, p\left(H_0; X\right) \leq \alpha^{\text{opt}}\right) \overset{\text{def}}{=} \inf_{\alpha \geq p(H_0; x)} \frac{\alpha}{\gamma} \end{aligned}$$

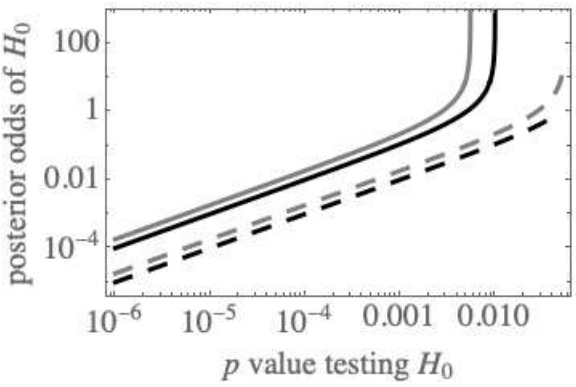


Figure 1: Upper bounds of the posterior odds that the data came from $H_0$, as a function of $p(H_0; x)$. $P_M(H_0)$ is 1/2 (black) or 10/11 (gray, following Benjamin et al. (2017)); $\gamma$ is $5 \times 10^{-3}$ (solid) or $5 \times 10^{-2}$ (dashed).

$$B_M(p(H_0; X) \leq p(H_0; x)) < \frac{1}{\gamma} p(H_0; x)$$
$$= B_M^{\text{upper}}\left(p(H_0; X) \leq \alpha^{\text{opt}}\right) \stackrel{\text{def}}{=} \inf_{\alpha \geq p(H_0; x)} \frac{\alpha}{\gamma},$$

where $\gamma$ satisfies Theorem 2's constraint (6).

For the first upper bound, Figure 1 displays $P_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}}) / (1 - P_M^{\text{upper}}(H_1 | p(H_0; X) \leq \alpha^{\text{opt}}))$ for different values of $p(H_0; x)$, $P_M(H_0)$, and $\gamma$. While the results certainly depend on the choices of $P_M(H_0)$ and $\gamma$, the overall message is clear: low $p$ values tend to lead to low posterior probabilities that the data came from the null hypothesis.

Not requiring the prior probability $P_M(H_0)$, either of the other two upper bounds, $R_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}})$ and $B_M^{\text{upper}}(p(H_0; X) \leq \alpha^{\text{opt}})$, might be reported in order 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, 2018b). However, Figure 2 suggests that $B_M^{\text{upper}}(p(H_0; X) \leq \alpha^{\text{opt}})$, the Bayes factor bound, is too conservative unless $P_M(H_0) \leq 1/2$.

For that reason, $R_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}})$, 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 $R_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}})$ by a hypothetical or estimated value of $P_M(H_0)$ that satisfies constraint (7). The resulting product is $P_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}})$, the optimal upper bound on the posterior probability.

If $P_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}})$ is sufficiently low, the null hypothesis may be considered false for decision

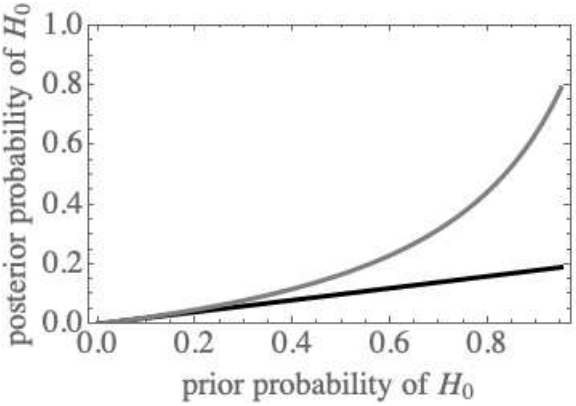


Figure 2: Upper bounds of the posterior probability of $H_0$ as a function of the prior probability $P_M(H_0)$ given $p(H_0; x) = 10^{-3}$ and $\gamma = 5 \times 10^{-3}$. The gray curve is derived from $B_M^{\text{upper}}(p(H_0; X) \leq \alpha^{\text{opt}})$ and Bayes's theorem; the black curve is $P_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}}) = P_M(H_0) R_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}})$.

making purposes since its posterior probability according to a model $M$ passing the check is even lower. When warranted, that can be formalized in terms of minimizing posterior expected loss or, considering $[0, P_M^{\text{upper}}(H_0 | p(H_0; X) \leq \alpha^{\text{opt}})[$ as an imprecise probability, in terms of a generalization of minimizing posterior expected loss (e.g., Troffaes, 2007).

**Example 4.** Example 3, continued. Let $M$ be any Bayesian model that passes the model check. Since constraint (7) is satisfied, Corollary 1 yields equation (8) and thus

$$P_M\left(H_0 | p(H_0; X) \lessapprox 5 \times 10^{-7}\right) \lessapprox \frac{10/11}{5 \times 10^{-3}} 5 \times 10^{-7} \approx 10^{-4},$$

indicating that the low $p$ value is not illusory. ▲

**Example 5.** Bernardo (2011) tested the null hypothesis $H_0$ that there is no measured effect due to extrasensory perception and that, in light of the sample with a size of about $10^6$, 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 $3 \times 10^{-4}$, $P_M(H_0 | X = x)$ would be over 95% if $P_M(H_0) = 1/2$ (Bernardo, 2011). By contrast, following the procedure of Example 4 with $\gamma = 5 \times 10^{-3}$, formula (8) yields

$$P_M\left(H_0 | p(H_0; X) \leq 3 \times 10^{-4}\right) \lessapprox \frac{5 \times 10^{-1}}{5 \times 10^{-3}} 3 \times 10^{-4} \approx 3 \times 10^{-2} = 0.03,$$

indicating a high posterior probability that there is some systematic error. This case obeys constraint (7):

$$P_M\left(H_0\right) < \frac{1-\gamma}{1-p\left(H_0;x\right)} = \frac{1-5\times10^{-3}}{1-3\times10^{-4}} = 99.5\%. \tag{9}$$

If the systematic bias could be ruled out, then $H_0$ would say there is no measured effect due to extrasensory perception, in which case $P_M\left(H_0\right)$ would be closer to 100% than is allowed by equation (9), and thus formula (8) would not apply. ▲

## Acknowledgments

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

## References

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., Johnson, V. E., 9 2017. Redefine statistical significance. Nature Human Behaviour, 1.

Bernardo, J. M., 2011. Integrated objective bayesian estimation and hypothesis testing. Bayesian statistics 9, 1–68.

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.

Bickel, D. R., 2018a. Confidence distributions and empirical Bayes posterior distributions unified as distributions of evidential support, working paper, DOI: 10.5281/zenodo.2529438.
URL https://doi.org/10.5281/zenodo.2529438

Bickel, D. R., 2018b. Reporting Bayes factors or probabilities to decision makers of unknown loss functions. Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2018.1459713.

Bickel, D. R., 2018c. Sharpen statistical significance: Evidence thresholds and Bayes factors sharpened into Occam's razors, working paper, HAL-01851322.
URL https://hal.archives-ouvertes.fr/hal-01851322

Cousins, R. D., Feb 2017. The jeffreys–lindley paradox and discovery criteria in high energy physics. Synthese 194, 395–432.

Evans, M., 2015. Measuring Statistical Evidence Using Relative Belief. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, New York.

Held, L., Ott, M., 2016. How the maximal evidence of p-values against point null hypotheses depends on sample size. American Statistician 70 (4), 335–341.

Held, L., Ott, M., 2018. On p-values and Bayes factors. Annual Review of Statistics and Its Application 5, 393–419.

Hill, J. R., 1990. A general framework for model-based statistics. Biometrika 77, 115–126.

Hjort, N., Dahl, F., Steinbakk, G., 2006. Post-processing posterior predictive p values. Journal of the American Statistical Association 101 (475), 1157–1174.

Lavine, M., Schervish, M. J., 1999. Bayes factors: What they are and what they are not. American Statistician 53, 119–122.

Lindley, D. V., 1957. A statistical paradox. Biometrika 44, pp. 187–192.

Micheas, A. C., Dey, D. K., 2003. Prior and posterior predictive p-values in the one-sided location parameter testing problem. Sankhya: The Indian Journal of Statistics (2003-) 65, 158–178.

O'Hagan, A., 2012. Higgs boson – digest and discussion. Technical Report, http://tonyohagan.co.uk/academic/pdf/HiggsBoson.pdf, accessed 31 December 2018.

Sellke, T., Bayarri, M. J., Berger, J. O., 2001. Calibration of p values for testing precise null hypotheses. American Statistician 55, 62–71.

Trafimow, D., 2003. Hypothesis testing and theory evaluation at the boundaries: Surprising insights from bayes's theorem. Psychological review 110 (3), 526.

Trafimow, D., Marks, M., 2015. Editorial. Basic and Applied Social Psychology 37 (1), 1–2.

Troffaes, M. C. M., 2007. Decision making under uncertainty using imprecise probabilities. International Journal of Approximate Reasoning 45 (1), 17–29.

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

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.

Zhao, G., Xu, X., 2014. The one-sided posterior predictive p-value for Fieller's problem. Statistics and Probability Letters 95, 57–62.