Sharpen statistical significance: Evidence thresholds and Bayes factors sharpened to Occam's razors

July 29, 2018

David R. Bickel

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

Abstract

Occam's razor suggests assigning more prior probability to a hypothesis corresponding to a simpler distribution of data than to a hypothesis with a more complex distribution of data, other things equal. An idealization of Occam's razor in terms of the entropy of the data distributions tends to favor the null hypothesis over the alternative hypothesis. As a result, lower p values are needed to attain the same level of evidence. A recently debated argument for lowering the significance level to 0.005 as the p value threshold for a new discovery and to 0.05 for a suggestive result would then support further lowering them to 0.001 and 0.01, respectively.

Keywords: Bayesian model averaging; Bayesian model selection; calibration of achieved error rates; empirical Bayes methods; foundations of statistics; hierarchical model; hypothesis testing; law of likelihood; likelihood paradigm; model checking; objective Bayes factor; p value calibration; strength of statistical evidence

1 Introduction

In their arguments to redefine statistical significance from 0.05 to 0.005, Benjamin et al. (2017) added to the "large and ever-increasing literature on the use and misuse of significance tests," with much of it centering on the fact that the p value is not defined as a probability that the null hypothesis is true (Cox, 2006, pp. 41-42), and the resulting difficulty in interpretation (Schervish, 1996; Royall, 1997; Efron and Gous, 2001; Goodman, 2003). The fact that p values can be automatically calculated makes them ubiquitous in scientific reports but also contributes to the perception that they are more objective than posterior probabilities.

Rather than choosing between significance testing and Bayesian methods, many researchers provided ways to calibrate a p value for interpretation as a Bayes factor, enabling interpretation as a posterior probability of null hypothesis truth when a prior probability can be specified. Such calibrations support not only the intuition of Fraser et al. (2004) and others that an extremely low p value indicates strong evidence against the null hypothesis but also support arguments against considering p values near 0.05 as indicative of strong evidence (e.g., Sellke et al., 2001). See Held and Ott (2018) for a recent review.

Under H_0, the null hypothesis, a p value $p(X)$ is a random variable with the uniform probability density function f_0, where X is the random vector that models the observable sample. With x as the observed sample, $p(x)$ then denotes the observed p value. Under the alternative hypothesis H_1, the probability density function of $p(X)$ is denoted by f_1. Thus, $p(X) \sim f = \pi_0 f_0 + \pi_1 f_1$, where $\pi_0 = P(H_0)$ and $\pi_1 = P(H_1)$ are the prior probabilities of the null and alternative hypotheses, respectively $(\pi_0 + \pi_1 = 1)$. $B(p(x)) = f_0(p(x)) / f_1(p(x))$ is called the *Bayes factor* in favor of the null hypothesis over the alternative hypothesis since

multiplying it by the prior odds yields the posterior odds, that is,

$$\frac{P\left(H_0|p\left(x\right)\right)}{P\left(H_1|p\left(x\right)\right)} = B\left(p\left(x\right)\right)\frac{\pi_0}{\pi_1}, \tag{1}$$

by Bayes's theorem,

$$P\left(H_0|p\left(x\right)\right) = \frac{\pi_0 f_0\left(p\left(x\right)\right)}{f\left(p\left(x\right)\right)}. \tag{2}$$

Guided by Occam's razor, Section 2 proposes a method of adjusting the prior probability π_0 for the simplicity of H_0 relative to that of H_1. That is accomplished indirectly by replacing an estimate of the Bayes factor such as $\widehat{B} \approx 1.65\left|z\left(x\right)\right|\exp\left(-\frac{z^2(x)}{2}\right)$ with a simplicity-adjusted estimate such as

$$\left|z\left(x\right)\right|\widehat{B}\left(p\left(x\right)\right) \approx 1.65z^2\left(x\right)e^{-\frac{z^2(x)}{2}}, \tag{3}$$

where $z\left(x\right)$ is a z-score corresponding to $p\left(x\right)$. For example, $z\left(x\right)$ could be the standard normal quantile of a one-sided p value testing whether a real-valued quantity is 0, or $z^2\left(x\right)$ could be the χ_1^2-quantile of a p value testing whether a non-negative quantity is 0, χ_1^2 being the χ^2 distribution with 1 degree of freedom.

Equation (3) may hold when $\left|z\left(x\right)\right| > 1$, indicating that simplicity considerations increase the Bayes factor and thus the evidence for the null hypothesis in proportion to $\left|z\left(x\right)\right|$, the extent that the p value is small. As a result, the p value has to be smaller when considering simplicity than otherwise to achieve a given level of evidence as quantified by a threshold of the Bayes factor. That is the basis of Section 3's case for redefining the p value threshold for statistical significance to 0.001 instead of the 0.005 level that Benjamin et al. (2017) advocate.

Rather than adjusting π_0 for the simplicity of H_0 relative to that of H_1, Section 4 adjusts a hyperprior distribution for the simplicity of the prior distribution, leading to similar results. The rationale for simplicity adjustments is clarified in the discussion, Section 5. The methods adjusted for hypothesis simplicity are extended in Appendix A to confidence-based methods of propagating uncertainty and in Appendix B to the case of simultaneously testing multiple null hypotheses.

2 Bayes factors adjusted for the simplicity of each hypothesis

2.1 How to adjust a Bayes factor for hypothesis simplicity

2.1.1 Likelihood functions adjusted for hypothesis simplicity

This subsection presents the general method of adjusting a likelihood function of a parameter for the simplicity of the distributions corresponding to the parameter values. Bickel (2016) proposed it as an aid to specifying prior distributions, quantifying simplicity in terms of Kolmogorov complexity.

If $g_\theta(y)$ is the probability density of an observed sample y given a parameter value θ, then $g_\theta(y)$ as a function of θ for a given y is called the *unsharpened likelihood function*. It is a likelihood function in the usual sense, and "unsharpened" indicates that it has not been sharpened, that is, adjusted to account for simplicity as per Occam's razor. Let $g_\theta^{(1)}(y_1)$ denote the probability density of y_1, a single value observed as a component of y. For each value of θ, the probability density function g_θ is a hypothesis about the distribution of Y, a random sample.

The simplicity of the hypothesis that $Y \sim g_\theta$ is the lack of complexity of $g_\theta^{(1)}$ in the sense of $S\left(g_\theta^{(1)}\right)$, the entropy of $g_\theta^{(1)}$. If, as is usual for a continuous y_1, $g_\theta^{(1)}$ is defined with respect to the Lebesgue measure, then the relevant complexity is the differential entropy,

$$S\left(g_\theta^{(1)}\right) = -\int g_\theta^{(1)}(y_1) \ln g_\theta^{(1)}(y_1)\, dy_1.$$

For an extended real number $\kappa \in [-\infty, \infty]$, the *$\kappa$-sharpened likelihood function* is

$$L_\kappa(\theta) = e^{-\kappa S\left(g_\theta^{(1)}\right)} g_\theta(y) \tag{4}$$

as a function of θ.

The *sharpness* κ controls the degree to which the unsharpened likelihood function is adjusted for hypothesis simplicity. At one extreme, L_0 is the unsharpened likelihood function, ignoring hypothesis simplicity, whereas at the other, L_∞ emphasizes simplicity to the extent of ignoring y. L_1 is the natural default, and L_2 is intermediate to L_1 and L_∞ inasmuch as 2 is the harmonic mean of 1 and ∞.

Bickel (2016) and Bickel (2018) originally considered $\kappa = 1$ and $\kappa \geq 0$, respectively. The case of $\kappa < 0$ is allowed as a way to implement Alexandre Patriota's suggestion to give more rather than less weight to distributions of higher entropy (private communication, May 24, 2018). Except when otherwise specified, $\kappa = 1$.

2.1.2 Bayes factors adjusted for hypothesis simplicity

The Bayes factor $B(p(x)) = f_0(p(x)) / f_1(p(x))$, not incorporating the simplicity of f_0 or f_1, is called the *unsharpened Bayes factor*. It may be sharpened to degree κ by setting $y = p(x)$ and $g_a = g_a^{(1)} = f_a$, where a, the indicator of the truth of the alternative hypothesis, satisfies

$a = 0$ under H_0 and $a = 1$ under H_1. In general, $f_a(p(x))$ is an integrated likelihood function of a since f_0 and f_1 are the prior predictive distributions of $p(X)$ under H_0 and H_1, respectively.

Thus, equation (4) gives $L_\kappa(a) = e^{-\kappa S\left(f_a^{(1)}\right)} f_a(p(x))$ as the κ-sharpened (integrated) likelihood function of a and

$$B_\kappa(p(x)) = \frac{L_\kappa(0)}{L_\kappa(1)} = \frac{e^{-\kappa S(f_0)} f_0(p(x))}{e^{-\kappa S(f_1)} f_1(p(x))} = e^{-\kappa\Delta} B(p(x)) \tag{5}$$

as the *κ-sharpened Bayes factor*, where $\Delta = S(f_0) - S(f_1)$. $B_\kappa(p(x))$ recovers the unsharpened Bayes factor when $\kappa = 0$ or $\Delta = 0$.

2.2 Example: The normal Bayes factor and its adjustment for hypothesis simplicity

For a running example of adjusting a Bayes factor for the simplicity of the null hypothesis and the alternative hypothesis, a Bayes factor based on the normal distribution will be used. While that is presented in Held and Ott (2016) as a lower bound on the Bayes factor, it is developed here instead as a maximum likelihood estimate of the Bayes factor in order to connect it with the fiducial methods of Appendix A and to clarify its relation to the empirical Bayes methods of multiple testing covered in Appendix B.

Let $p_{\text{one}}(x)$ denote a given one-sided p value and $p(x) = 2\min(p_{\text{one}}(x), 1 - p_{\text{one}}(x))$ the corresponding two-sided p value. Alternatively, if $p(x)$ is the p value for a χ^2 test or for another test of a null hypothesis at the boundary of the parameter space, then let $p_{\text{one}}(x) = 1 - p(x)/2$. In either case, $z(x)$ denotes the $p_{\text{one}}(x)$-quantile for the standard normal distribution function Φ, that is, $z(x) = \Phi^{-1}(p_{\text{one}}(x))$.

Under H_0, $z(X) \sim \mathrm{N}(0,1)$ since $p(X) \sim \mathrm{U}(0,1)$. When assuming $z(X) \sim \mathrm{N}(0,\sigma^2)$ for some $\sigma > 1$ under H_1, the maximum likelihood estimate of $B(p(x))$ is the *normal Bayes factor*,

$$\widehat{B}(p(x)) = \frac{\phi(z(x))}{\phi(\widehat{\sigma}(x) z(x))} = \frac{\phi(z(x))}{\phi(\widehat{\sigma}(x) z(x))} = \widehat{\sigma}(x) e^{-\frac{\left(1-\widehat{\sigma}^{-2}(x)\right) z^2(x)}{2}}$$

for $|z(x)| \geq 1$, where ϕ is the standard normal density function, and $\widehat{\sigma}(x)$ is defined as σ's maximum likelihood estimate,

$$\widehat{\sigma}(x) = \arg\sup_{\sigma>0} \phi(\sigma z(x)) = |z(x)|.$$

Thus, the normal Bayes factor is

$$\widehat{B}(p(x)) = |z(x)| e^{-\frac{z^2(x)-1}{2}} = \sqrt{e}\,|z(x)| e^{-\frac{z^2(x)}{2}} \approx 1.65\,|z(x)| e^{-\frac{z^2(x)}{2}}. \tag{6}$$

Since $\widehat{B}(p(x))$ is not adjusted for hypothesis simplicity, it is unsharpened in the sense that it is an estimate of $B(p(x))$, a unsharpened Bayes factor in the terminology of Section 2.1. Equation (5) suggests the *sharpened normal Bayes factor*,

$$\widehat{B}_\kappa(p(x)) = e^{-\kappa\widehat{\Delta}} \widehat{B}(p(x)),$$

where $\widehat{\Delta} = H(\mathrm{N}(0,1)) - H(\mathrm{N}(0,\widehat{\sigma}^2))$. From Michalowicz et al. (2013, p. 127), $\widehat{\Delta} = \ln 1 - \ln\widehat{\sigma} = -\ln\widehat{\sigma}$, leading to $\widehat{B}_\kappa(p(x)) = \widehat{\sigma}^\kappa \widehat{B}(p(x)) = |z(x)|^\kappa \widehat{B}(p(x))$. The $\widehat{B}_1(p(x))$ case appears in equation (3).

2.3 Comparisons to other Bayes factors

To compare the sharpened Bayes factor $\widehat{B}_1(p(x))$ to its unsharpened counterpart $\widehat{B}_0(p(x))$ (§2.2) and to the following lower bounds of the Bayes factor, Figure 1 displays the quantities as functions of $p(x)$. There, the *universal lower bound on the Bayes factor* is $\underline{\underline{B}}(p(x)) = e^{-z^2(x)}$ (Held and Ott, 2016).

What Figure 1 calls the *conservative lower bound on the Bayes factor* is

$$\underline{B}(p(x)) = -ep(x)\ln p(x), \tag{7}$$

which Sellke et al. (2001) considered as a lower bound on $B(p(x))$ under $p(x) \leq 1/e$ and a broad condition on the hazard rate that is useful for testing simple, two-sided null hypotheses. As Sellke et al. (2001) lamented, an upper bound $\overline{B}(p(x))$ would be more desirable than a lower bound since a sufficiently low value of the upper bound would guarantee any specified amount of evidence against H_0, the same not being true of a lower bound. For example, $\overline{B}(p(x)) < 1\%$ implies that $B(p(x)) < 1\%$, but $\underline{B}(p(x)) < 1\%$ is compatible with $B(p(x)) > 1\%$. However, $\underline{B}(p(x))$ is nonetheless higher than (more conservative than) $\underline{\underline{B}}(p(x))$ and the other lower bounds of Bayes factors plotted in Benjamin et al. (2017). Further, its proximity to $\widehat{B}_0(p(x))$, as seen in Figure 1, also suggests that it is not too low as an estimate of the unsharpened Bayes factor.

2.4 Posterior probability adjusted for the simplicity of each hypothesis

The posterior probability of H_0 may be calculated from its prior probability π_0 and the Bayes factor $B(p(x))$:

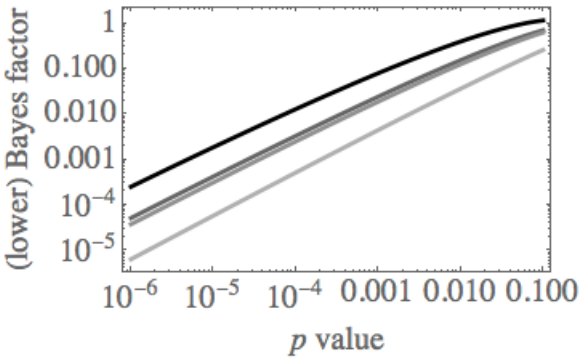

Figure 1: The estimated likelihood ratio or lower bound on the Bayes factor as a function of the two-sided p value. From highest to lowest and from darkest to lightest, the curves are the sharpened normal Bayes factor, the unsharpened normal Bayes factor, the conservative lower bound on the Bayes factor, and the universal lower bound on the Bayes factor. In symbols, they are $\widehat{B}_1(p(x))$, $\widehat{B}_0(p(x))$, $\underline{B}(p(x))$, and $\underline{\underline{B}}(p(x))$, respectively.

$$P(H_0|p(x)) = \left(1 + \left(\frac{P(H_0|p(x))}{P(H_1|p(x))}\right)^{-1}\right)^{-1} = \left(1 + \left(B(p(x))\frac{\pi_0}{1-\pi_0}\right)^{-1}\right)^{-1} \tag{8}$$

according to equations (1) and (2). Likewise, given π_0 and the κ-sharpened Bayes factor $B_\kappa(p(x))$, the *κ-sharpened posterior probability* of H_0 is

$$P_\kappa(H_0|p(x)) = \left(1 + \left(e^{-\kappa\Delta}B(p(x))\frac{\pi_0}{1-\pi_0}\right)^{-1}\right)^{-1} = \left(1 + \left(B(p(x))\frac{e^{-\kappa S(f_0)}\pi_0}{e^{-\kappa S(f_1)}(1-\pi_0)}\right)^{-1}\right)^{-1} \tag{9}$$

since equation (5) has $B_\kappa(p(x)) = e^{-\kappa\Delta}B(p(x))$ and since $\Delta = S(f_0) - S(f_1)$. The factor in the right-hand-side that is multiplied by $B(p(x))$ clarifies the rationale for sharpening the Bayes factor as an elegant way to apply Occam's razor to the assessment of the prior distribution, as will be seen in Section 5.

Negligible	Weak	Moderate	Strong	Very Strong	Overwhelming
$\widehat{W}_1 \geq 0$	$\widehat{W}_1 \geq 1$	$\widehat{W}_1 \geq 2$	$\widehat{W}_1 \geq 3$	$\widehat{W}_1 \geq 5$	$\widehat{W}_1 \geq 7$
$p \lessapprox 0.1$	$\boldsymbol{p \lessapprox 0.01}$	$p \lessapprox 0.005$	$\boldsymbol{p \lessapprox 0.001}$	$p \lessapprox 0.0005$	$p \lessapprox 0.00005$

Table 1: Scales of evidence for H_1 over H_0, with the first row according to intervals of the 1-sharpened weight of evidence in bits. The scale is an adaptation (Bickel, 2011) of the classic base-10 scales of Jeffreys (1948) to $W(p(x)) \geq 3$ and $W(p(x)) \geq 5$, the two base-2 scales of Royall (1997). Here, $\widehat{W}_1$ abbreviates $\widehat{W}_1(p(x))$, and p abbreviates $p(x)$. Two of the p value thresholds values are in boldface since they indicate when a result is suggestive ($\alpha = 0.01$) or significant ($\alpha = 0.001$) according to Section 3.

2.5 Strength of statistical evidence sharpened for hypothesis simplicity

To measure the strength of statistical evidence that the alternative hypothesis is true, let $W(p(x)) = -\log_2 B(p(x))$, $W_\kappa(p(x)) = -\log_2 B_\kappa(p(x))$, and $\widehat{W}_\kappa(p(x)) = -\log_2 \widehat{B}_\kappa(p(x))$. Adding "unsharpened" to the term from Good (1979), $W(p(x))$ is the *unsharpened weight of evidence* in the data favoring H_1 over H_0; analogously, $W_\kappa(p(x))$ and $\widehat{W}_\kappa(p(x))$ are the *κ-sharpened weight of evidence* and the *κ-sharpened normal weight of evidence*, respectively. $W_\kappa(p(x))$ reduces to the unsharpened weight of evidence if $\kappa = 0$ or if the hypotheses are equally simple in the distributional sense, that is, if $H(f_0) = H(f_1)$.

The weight of evidence is traditionally interpreted in terms of grades of evidence like those found in Table 1, with $W(p(x))$ in place of its $\widehat{W}_1$. The p value thresholds shown there are based on $\widehat{W}_1(p(x))$, the 1-sharpened normal weight of evidence.

3 Statistical significance adjusted for the simplicity of each hypothesis

Benjamin et al. (2017) argued that 0.005 should be the p value threshold for saying that a result is significant enough to claim a discovery and that the 0.05 threshold should instead indicate that a result is suggestive. The argument is based largely on the Bayes factors that would be attained at each of those thresholds. Those Bayes factors were not adjusted for the simplicity of each hypothesis.

Since the simplicity-adjusted Bayes factors are higher than those not adjusted, the p value thresholds needed to attain the same Bayes factor values are lower when simplicity is taken into account. Sharpening the normal Bayes factor for simplicity leads to these adjustments in the p value thresholds:

- The statistical significance threshold of 0.005 (Benjamin et al., 2017) corresponds to a threshold of the unsharpened Bayes factor equal to 3.5. The sharpened Bayes factor is about 3.5 when the p value is 0.001, yielding 0.001 as the simplicity-adjusted threshold for a significant result. As that is Table 1's threshold for strong evidence, that means statistical significance would only be declared when there is strong evidence against the null hypothesis.

- For a suggestive result, the p value threshold of 0.05 (Benjamin et al., 2017) corresponds to a threshold of the unsharpened Bayes factor equal to 1.1. The sharpened Bayes factor is about 1.1 when the p value is 0.01, yielding 0.01 as the simplicity-adjusted threshold for a suggestive result. That being Table 1's threshold for weak evidence, a result would only be called suggestive when the evidence against the null hypothesis

	$p(x) \leq 0.001$	$p(x) \leq 0.005$	$p(x) \leq 0.01$	$p(x) \leq 0.05$
Report	Significant if $\kappa = 1$	Significant if $\kappa = 0$	Suggestive if $\kappa = 1$	Suggestive if $\kappa = 0$
$\boldsymbol{\kappa = 0}$	$W_0(x) \geq 5.4$	$W_0(x) \geq 3.5^*$	$W_0(x) \geq 2.7$	$W_0(x) \geq 1.1^\dagger$
$\boldsymbol{\kappa = 1}$	$W_1(x) \geq 3.7^*$	$W_1(x) \geq 2.0$	$W_1(x) \geq 1.3^\dagger$	$W_1(x) \geq 0.1$
$\boldsymbol{\kappa = 2}$	$W_2(x) \geq 1.9$	$W_2(x) \geq 0.5$	$W_2(x) \geq 0.0$	$W_2(x) \geq -0.9$

Table 2: Weights of evidence corresponding to the 0.001 and 0.005 p value thresholds for a significant result and the 0.01 and 0.05 p value thresholds for a suggestive result. The higher of each threshold (0.005 or 0.05) is that proposed by Benjamin et al. (2017), and the lower (0.001 or 0.01) is what it would take for a Bayes factor sharpened at level $\kappa = 1$ to attain approximately the same weight of evidence ($3.5 \approx 3.7$ or $1.1 \approx 1.3$), as matched by $*$ for significant results and † for suggestive results. Here, $W_\kappa(x)$ is the base-2 logarithm of the Bayes factor in favor of the alternative hypothesis over the null hypothesis when the degree of sharpness is κ.

is weak but not negligible.

Table 2 summarizes that method of adjusting the p value thresholds for the simplicity of the null hypothesis and alternative hypothesis using $\kappa = 1$ as the amount of the adjustment, with $\kappa = 0$ corresponding to no simplicity adjustment. (The weights of evidence for $\kappa = 2$ are also shown in the table since that is an intermediate value in the sense that 2 is the harmonic mean of 1 and ∞.)

4 Posterior probability adjusted for the simplicity of each π_0 value

In Section 2, the prior probability that the null hypothesis is true was assessed as guided by the simplicity of H_0's prior predictive distribution compared to the simplicity of H_1's prior predictive distribution. Using π_0 as the prior probability without simplicity considerations, the simplicity adjustment to the prior was passed to the Bayes factor for the reasons to be

discussed in Section 5. In contrast with thereby treating π_0 as an epistemological probability in need of assessment with help from Occam's razor, this section treats π_0 as an unknown probability that is a property of the system studied. That means the simplicity of π_0 can guide the assessment of the its higher-level prior distribution, a hyperprior distribution.

The alternative hypothesis indicator A is defined as a random variable of Bernoulli distribution $\text{Bern}(1-\pi_0)$, that is, $P(A=0)=\pi_0$ and $P(A=1)=1-\pi_0$. That A is equal to 1 if H_1 is true and that is equal to 0 if H_0 is true.

Just as Section 2.1.1 quantified the simplicity of each hypothesis in terms of its entropy of $p(X)$, this section quantifies the simplicity of each value of π_0 in terms of its entropy of A. The general method of Section 2.1.1 applies to that entropy and hierarchical model with $Y=A$, $\theta=\pi_0$, and $g_\theta=g_{\pi_0}=\text{Bern}(1-\pi_0)$, as follows. Let Π denote the unsharpened hyperprior distribution of π_0. If Π is a probability density function with respect to the Lebesgue measure or a probability mass function, then the *κ-sharpened hyperprior probability density* or the *κ-sharpened hyperprior probability mass* of π_0 is $\Pi_\kappa(\pi_0) \propto e^{-\kappa S\left(g_{\pi_0}\right)}\Pi(\pi_0)$, that is,

$$\Pi_\kappa(\pi_0)=\frac{e^{-\kappa S\left(g_{\pi_0}\right)}\Pi(\pi_0)}{\int e^{-\kappa S\left(g_{\pi_0'}\right)}\Pi(\pi_0')\,d\pi_0'} \text{ or } \Pi_\kappa(\pi_0)=\frac{e^{-\kappa S\left(g_{\pi_0}\right)}\Pi(\pi_0)}{\sum_{\pi_0'} e^{-\kappa S\left(g_{\pi_0'}\right)}\Pi(\pi_0')} \tag{10}$$

respectively. The entropy of $\text{Bern}(1-\pi_0)$ is

$$\begin{aligned} S(g_{\pi_0}) &= (-\pi_0 \ln \pi_0 - (1-\pi_0)\ln(1-\pi_0)) \\ &= -\ln\left(\pi_0^{\pi_0}(1-\pi_0)^{(1-\pi_0)}\right); \end{aligned}$$

$$\therefore \Pi_\kappa(\pi_0) \propto e^{-\kappa S\left(g_{\pi_0}\right)}\Pi(\pi_0) \propto \pi_0^{\pi_0}(1-\pi_0)^{(1-\pi_0)}.$$

The corresponding *κ-sharpened hyperposterior probability density* or the *κ-sharpened hy-*

perposterior probability mass is then given by

$$\Pi_\kappa\left(\pi_0|p\left(x\right)\right) \propto \Pi_\kappa\left(\pi_0\right)\left(\pi_0 f_0\left(p\left(x\right)\right) + \left(1-\pi_0\right) f_1\left(p\left(x\right)\right)\right).$$

The *κ-sharpened hyperposterior probability probability* that H_0 is true is, as per equation (8),

$$\begin{aligned} P_{\Pi_\kappa}\left(H_0|p\left(x\right)\right) &= \sum_{\pi_0} \Pi_\kappa\left(\pi_0|p\left(x\right)\right) \frac{\pi_0 f_0\left(p\left(x\right)\right)}{\pi_0 f_0\left(p\left(x\right)\right) + \left(1-\pi_0\right) f_1\left(p\left(x\right)\right)} \\ &= \sum_{\pi_0} \Pi_\kappa\left(\pi_0|p\left(x\right)\right) \left(1 + \left(B\left(p\left(x\right)\right) \frac{\pi_0}{1-\pi_0}\right)^{-1}\right)^{-1} \end{aligned} \tag{11}$$

in the probability mass case and the analogous quantity in the probability density case.

Example. Suppose, as in Section 2.2, that $f_0 = \mathrm{N}\left(0,1\right)$ and $f_1 = \mathrm{N}\left(0,\sigma^2\right)$, and let Π be the probability mass function on the domain $\{1/2, 1\}$ of two possible values of π_0 such that $\Pi_\kappa\left(1/2\right) = \Pi_\kappa\left(1\right) = 1/2$. To compare the resulting adjustment for π_0 simplicity to the corresponding adjustment for hypothesis simplicity (§2), Figure 2 displays two proposed sharpened posterior probability estimates and their unsharpened counterparts as functions of $p\left(x\right)$, the two-sided p value.

5 Discussion

This paper draws out implications of sharpening Bayes factors and the p value thresholds and posterior probabilities that depend on such Bayes factors, where "sharpening" means adjusting them for the simplicity of distributions as motivated by Occam's razor. But why should they be adjusted for simplicity? Even if Occam's razor is applicable to distributional simplicity, why sharpen Bayes factors as opposed to prior probabilities?

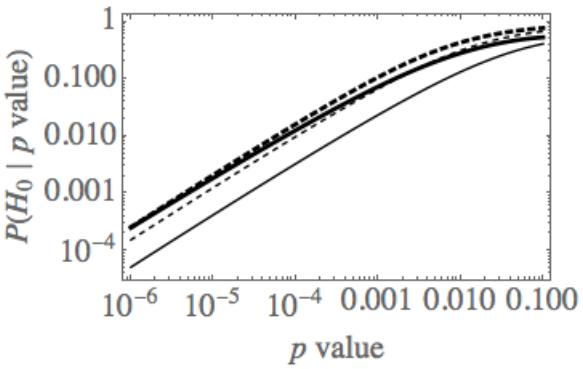

Figure 2: Estimated posterior probability that the null hypothesis is true as a function of the two-sided p value. The solid curves are based on $\pi_0 = 1/2$ (50% unsharpened prior probability that the null hypothesis is true); the dashed curves are based on a 50% unsharpened hyperprior probability that $\pi_0 = 1/2$ and 50% that $\pi_0 = 1$; the thicker curves represent the sharpened versions of the thinner curves. Thus, in the notation of equations (9) and (11), they are $P_1\left(H_0|p\left(x\right)\right)$ (solid, thicker), $P_0\left(H_0|p\left(x\right)\right)$ (solid, thinner), $P_{\Pi_1}\left(H_0|p\left(x\right)\right)$ (dashed, thicker), and $P_{\Pi_0}\left(H_0|p\left(x\right)\right)$ (dashed, thinner).

The second factor in the last term of equation (9) sheds light on the rationale for sharpening (adjusting for simplicity), for that factor may be interpreted as the sharpened prior odds corresponding to the *κ-sharpened prior probability*

$$\pi_{0,\kappa} = \frac{e^{-\kappa\, S(f_0)}\pi_0}{e^{-\kappa\, S(f_0)}\pi_0 + e^{-\kappa\, S(f_1)}\left(1-\pi_0\right)},$$

where $\kappa > 0$ is the degree of sharpening according to Occam's razor; $-\,S\left(f_0\right)$ and $-\,S\left(f_1\right)$ are the degrees of simplicity of the distribution of the p value under the null hypothesis and the alternative hypothesis, respectively. The motive for sharpening is that the simplicity of f_0 compared to that of f_1 may guide the specification of the prior probability. According to this rationale, π_0 is the unsharpened prior probability of H_0, that is, the prior probability that would be assigned without consideration of its distributional simplicity relative to that of H_1. Then $\pi_{0,\kappa}$ is a better considered prior probability in the sense that it accounts for

that relative simplicity (Bickel, 2016, 2018).

Using the sharpened prior with the unsharpened Bayes factor yields the same posterior probability as using the unsharpened prior with the sharpened Bayes factor, as may be seen in equation (9). Whereas the rationale says an unsharpened prior should be sharpened, arranging the mathematics for instead sharpening the Bayes factor enables much clearer comparisons with previous Bayes factors without having to consider the unsharpened prior, as in Sections 2 and 3.

That rationale applies only at a higher level to Section 4, in which π_0 is an unknown parameter of the biological system or other system studied rather than an epistemological probability. There, the epistemological probability is modeled as a hyperprior.

Acknowledgments

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

A Replacing maximum likelihood with confidence-based methods

A.1 A fiducial modification of the normal Bayes factor

Let $\sigma_q = |z(x)| / \sqrt{F^{-1}(q)}$, where $F^{-1}(q)$ is the qth quantile of a χ^2_1 variate. Without knowledge of the Bayes factor $B(p(x))$, this counterpart may be calculated for $|z(x)| \geq 1$

and a $q \in [0,1]$:

$$B\left(p\left(x\right);q\right)=\frac{\phi\left(z\left(x\right)\right)}{\phi\left(\sigma_{q}z\left(x\right)\right)}=\frac{\phi\left(z\left(x\right)\right)}{\phi\left(\sigma_{q}z\left(x\right)\right)}=\sigma_{q}e^{-\frac{\left(1-\sigma_{q}^{-2}\right)z^{2}(x)}{2}}=\frac{\left|z\left(x\right)\right|}{\sqrt{F^{-1}\left(q\right)}}e^{-\frac{z^{2}(x)-F^{-1}(q)}{2}} \quad (12)$$

in terms of the notation of Section 2.2.

To interpret $B\left(p\left(x\right);q\right)$, consider the random variable $Q \sim \mathrm{U}\left(0,1\right)$. For each $q \in [0,1]$, the value $B\left(p\left(x\right);q\right)$ is the qth quantile of $B\left(p\left(x\right);Q\right)$, a random variable with a fiducial distribution that depends on $p\left(x\right)$, the fixed p value (Bickel, 2017).

Thus, $B\left(p\left(x\right)\right)$ may be estimated by $B\left(p\left(x\right);Q\right)$'s median (Bickel, 2017),

$$B\left(p\left(x\right);1/2\right)=\frac{\phi\left(z\left(x\right)\right)}{\phi\left(\sigma_{1/2}z\left(x\right)\right)}=\sigma_{1/2}e^{-\frac{\left(1-\sigma_{1/2}^{-2}\right)z^{2}(x)}{2}}.$$

Since $F^{-1}\left(1l2\right)$ is the median of a χ_{1}^{2} variate, it follows that $\sigma_{1/2} \approx 1.48\left|z\left(x\right)\right|$ and

$$B\left(p\left(x\right);1/2\right) \approx 1.86\left|z\left(x\right)\right|e^{-\frac{z^{2}(x)}{2}},$$

which equation (6) shows to be essentially the same as the maximum likelihood estimate.

A.2 Fiducial estimates of the posterior probability of H_0

A fiducial estimate of the posterior probability of the null hypothesis may be found by substituting $B\left(p\left(x\right);1/2\right)$ for $B\left(p\left(x\right)\right)$ in equation (9). Bickel (2017) considered the unsharpened ($\kappa = 0$) versions of that and other ways to use confidence distributions and other coherent fiducial distributions to propagate the uncertainty in posterior probabilities.

Those methods are equally applicable to sharpened posterior probabilities. For example,

a κ-sharpened fiducial posterior probability of H_0 is the expectation value

$$E_{Q\sim \mathrm{U}(0,1)}\left(1+\left(B\left(p\left(x\right);Q\right)\frac{e^{-\kappa\, S(f_0)}\pi_0}{e^{-\kappa\, S(f_1)}\left(1-\pi_0\right)}\right)^{-1}\right)^{-1}.$$

Appendix B bridges the gap between the empirical Bayes terminology of Bickel (2017) and the evidential terminology of this paper.

B Multiple testing and posterior probabilities as local false discovery rates

From an empirical Bayes point of view, equation (2)'s $P\left(H_0|p\left(x\right)\right)$ is an unknown *local false discovery rate* (Bickel, 2014, 2017), so named because of its relation to the false discovery rate in the context of testing multiple hypotheses, replacing the $p\left(x\right)$ with a vector of p values, one for each hypothesis tested (Efron, 2010). In empirical Bayes terminology, the $P_\kappa\left(H_0|p\left(x\right)\right)$ of equation (9) would then be the *κ-sharpened local false discovery rate*.

In the setting of m simultaneous hypothesis tests, the p values are denoted by $z\left(x_1\right)$, $z\left(x_2\right)$, $\ldots$, $z\left(x_m\right)$, where each x_i represents a different sample. As part of an empirical Bayes approach to hierarchical models, Efron (2007, 2010, pp. 72-74) considered normal distributions of $z\left(X\right)$ conditional on the alternative hypothesis H_1, the proposition that the null hypothesis $H_0 : z\left(X\right) \sim \mathrm{N}\left(0,1\right)$ is false. In short, the alternative hypothesis is that $H_1 : z\left(X\right) \sim \mathrm{N}\left(\mu_1,\sigma\right)$ with $\left(\mu_1,\sigma\right) \neq \left(0,1\right)$, that is, with $\mu_1 \neq 0$ and/or $\sigma \neq 1$. In his terminology, H_0 is the "theoretical null hypothesis" since its distribution of $z\left(X\right)$ does not depend on x, the observation.

In the example of Section 2.2, the lack of background information requires the alternative

hypothesis to be so vague that $\mu = 0$, making H_1 an example of a local alternative hypothesis (Held and Ott, 2018), and $\sigma > 1$. Equation (6) then generalizes to

$$\widehat{B}\left(p\left(x_i\right); m\right) = \widehat{\sigma} e^{-\frac{\left(1-\widehat{\sigma}_1^{-2}\right) z^2(x)}{2}},$$

where $\widehat{\sigma}$ is the maximum likelihood estimate of the standard deviation of $z\left(x_1\right)$, $z\left(x_2\right)$, ..., $z\left(x_m\right)$, with each $z\left(x_i\right)$ related to $p\left(x_i\right)$ in the same way as Section 2.2 related $z\left(x\right)$ to $p\left(x\right)$. (The case $\widehat{B}\left(p\left(x_1\right); 1\right)$ then recovers $\widehat{B}\left(p\left(x_1\right)\right)$.) An empirical Bayes estimate of the local false discovery rate of the ith null hypothesis is obtained by plugging $\widehat{B}\left(p\left(x_i\right); m\right)$ into $B\left(p\left(x\right)\right)$ in equation (9).

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.

Bickel, D. R., 2011. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. Canadian Journal of Statistics 39, 610–631.

Bickel, D. R., 2014. Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison. International Statistical Review 82, 457–476.

Bickel, D. R., 2016. Computable priors sharpened into Occam's razors, working paper, HAL-01423673.
URL https://hal.archives-ouvertes.fr/hal-01423673

Bickel, D. R., 2017. Confidence distributions applied to propagating uncertainty to inference based on estimating the local false discovery rate: A fiducial continuum from confidence sets to empirical Bayes set estimates as the number of comparisons increases. Communications in Statistics - Theory and Methods 46 (21), 10788–10799.

Bickel, D. R., 2018. Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors, working paper, HAL-01799519.
URL https://hal.archives-ouvertes.fr/hal-01799519

Cox, D. R., 2006. Principles of Statistical Inference. Cambridge University Press, Cambridge.

Efron, B., 2007. Size, power and false discovery rates. Annals of Statistics 35, 1351–1377.

Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press, Cambridge.

Efron, B., Gous, A., 2001. Scales of evidence for model selection: Fisher versus Jeffreys. Lecture Notes - Monograph Series 38, 208–256.

Fraser, D. A. S., Reid, N., Wong, A. C. M., 2004. Inference for bounded parameters. Physical Review D 69, 033002.

Good, I. J., 1979. Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II. Biometrika 66 (2), 393–396.

Goodman, S., 2003. Commentary: The p-value, devalued. International Journal of Epidemiology 32, 699–702.

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.

Jeffreys, H., 1948. Theory of Probability. Oxford University Press, London.

Michalowicz, J. V., Nichols, J. M., Bucholtz, F., 2013. Handbook of Differential Entropy. CRC Press, New York.

Royall, R., 1997. Statistical Evidence: A Likelihood Paradigm. CRC Press, New York.

Schervish, M. J., 1996. P values: What they are and what they are not. American Statistician 50, 203–206.

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