Interval estimation, point estimation, and null hypothesis significance testing calibrated by an estimated posterior probability of the null hypothesis

David R. Bickel

Interval estimation, point estimation, and null hypothesis significance testing calibrated by an estimated posterior probability of the null hypothesis

March 2, 2020

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

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

dbickel@uottawa.ca

Abstract

Much of the blame for failed attempts to replicate reports of scientific findings has been placed on ubiquitous and persistent misinterpretations of the p value. An increasingly popular solution is to transform a two-sided p value to a lower bound on a Bayes factor. Another solution is to interpret a one-sided p value as an approximate posterior probability.

Combining the two solutions results in confidence intervals that are calibrated by an estimate of the posterior probability that the null hypothesis is true. The combination also provides a point estimate that is covered by the calibrated confidence interval at every level of confidence. Finally, the combination of solutions generates a two-sided p value that is calibrated by the estimate of the posterior probability of the null hypothesis. In the special case of a 50% prior probability of the null hypothesis and a simple lower bound on the Bayes factor, the calibrated two-sided p value is about (1 - abs(2.7 p ln p)) p + 2 abs(2.7 p ln p) for small p. The calibrations of confidence intervals, point estimates, and p values are proposed in an empirical Bayes framework without requiring multiple comparisons.

Keywords: calibrated effect size estimation; calibrated confidence interval; calibrated *p* value; replication crisis; reproducibility crisis

1 Introduction

The widespread failure of attempts to reproduce scientific results is largely attributed to the apparently incurable epidemic of misinterpreting p values. Much of the resulting debate about whether and how to test null hypotheses stems from the wide spectrum of attitudes toward frequentist and Bayesian schools of statistics, as seen in the special issue of *The American Statistician* introduced by (Wasserstein et al., 2019). For that reason, reconsidering the old probability interpretations that lurk beneath the expressed opinions may lead to new insights for resolving the controversy.

The subjective interpretation of the prior probability provides guidance in the selection of prior distributions even in practical situations in which it is not usually feasible to elicit the prior probability of any expert or other agent. In the case of reliable information about the value of the parameter, the subjective interpretation serves as a benchmark that gives meaning to assigning prior probabilities in the sense that they are interpreted as levels of belief that a hypothetical agent would have in the corresponding hypothesis (Bernardo, 1997) or as an approximation of the agent's levels of belief. Likewise, in the absence of reliable information about the value of the parameter, a default or reference prior is used either to determine the posterior probability distribution of a hypothetical agent whose beliefs distribution matches the reference prior distribution or as an approximation of the belief distribution of one or more individuals. Such priors are sometimes called "objective" in the sense that they are not purely individualistic since whole communities can agree on their use.

That motivation for objective priors contrasts with the logical-probability position popularized by Jaynes (2003) and called "objective Bayesianism" in the philosophical literature (Williamson, 2010). This school follows the students Jeffreys (1948) and Keynes (1921) of W. E. Johnson in insisting on less subjective foundations of Bayesianism but has found little favor in the statistical community due in part to the marginalization paradox (Dawid et al., 1973) and to the failure of decades of research to find any genuinely noninformative priors (Kass and Wasserman, 1996). While there are important differences between this school and that of traditional Bayesian statistics, probability is viewed by both as a level of belief, whether of a real agent or of a hypothetical, perfectly logical agent. Indeed, the difference is chiefly that of emphasis since the most influential subjective Bayesians used probability to prescribe coherent behavior rather than to describe the actual behavior of any human agent (Levi, 2008). (Carnap (1962, pp. 42-47; 1971, p. 13; 1980, p. 119) observed that, like H. Jeffreys, subjectivists such as F. P. Ramsay, B. de Finetti, and L. J. Savage sought normative theories of rational credence or decision, not ways to describe aspects

of anyone's actual state of mind or behavior, in spite of differences in emphases and misleading uses of psychologizing figures of speech, which not even R. A. Fisher managed to avoid entirely (Zabell, 1992).) Belief prescription, not belief description, is likewise sought in traditional Bayesian statistics (e.g., Bernardo and Smith, 1994).

All of those Bayesian interpretations of prior and posterior probability fall under the broad category of *belief-type probability*, which includes epistemic and evidential probabilities (e.g., Kyburg and Teng, 2001) not associated with any actual beliefs (Hacking, 2001, Ch. 12). Frequentist statisticians instead prefer to work exclusively with what Hacking (2001, Ch. 12) calls *frequency-type probability*, some kind of limiting relative frequency or propensity that is a feature of the domain under investigation, regardless of what evidence is available to the investigators. In short, belief-type probability refers to real or hypothetical states of knowledge, whereas frequency-type probability refers to real or hypothetical frequencies of events.

Seeking the objective analysis of data, frequentists avoid belief-type priors: "The sampling theory approach is concerned with the relation between the data and the external world, however idealized the representation. The existence of an elegant theory of self-consistent private behaviour seems no grounds for changing the whole focus of discussion" (Cox, 1978). Many wary of using Bayes's theorem with belief-type priors would nonetheless use it with frequency-type distributions of parameter values (Neyman, 1957; Fisher, 1973; Wilkinson, 1977; Edwards, 1992; Kyburg and Teng, 2001; Kyburg and Teng, 2006; Hald, 2007, p. 36; Fraser, 2011). In statistics, the random parameters in empirical Bayes models, mixed-effects models, and latent variable models have frequency-type distributions, representing variability rather than uncertainty. This viewpoint leads naturally to the empirical Bayes approach of estimating a frequency-type prior from data (Efron, 2008); see Neyman (1957)'s enthusiasm for early empirical Bayes work (Robbins, 1956).

Example 1. Consider the posterior probability of the null hypothesis conditional on p, the observed two-sided p value:

$$\Pr\left(\vartheta = \theta_{H_0} \mid P = p\right) = \frac{\Pr\left(\vartheta = \theta_{H_0}\right) \Pr\left(P = p \mid \vartheta = \theta_{H_0}\right)}{\Pr\left(P = p\right)} = \left(1 + \left(\frac{\Pr\left(\vartheta = \theta_{H_0}\right)}{1 - \Pr\left(\vartheta = \theta_{H_0}\right)} B\right)^{-1}\right)^{-1} \tag{1}$$

where ϑ and P are the parameter of interest and the p value as random variables, θ_{H_0} is the parameter value under the null hypothesis, the null hypothesis that $\vartheta = \theta_{H_0}$ asserts the truth of the null hypothesis to a sufficiently close approximation for practical purposes, and B is the Bayes factor $f\left(p \mid \theta = \theta_{H_0}\right) / f\left(p \mid \theta \neq \theta_{H_0}\right)$ based on f, the probability density function of P. When

$\Pr(\vartheta = \theta_{H_0} \,|\, P = p)$ is a frequency-type probability, it is known as the *local false discovery rate* (LFDR) in the empirical Bayes literature on testing multiple hypotheses (e.g., Efron et al., 2001; Efron, 2010) and on testing a single hypothesis (e.g., Bickel, 2017, 2019c,e). If the assumptions of Sellke et al. (2001) hold for frequency-type probability distributions underlying LFDR, then its lower bound is

$$\underline{\text{LFDR}} = \left(1 + \left(\frac{\Pr(\vartheta = \theta_{H_0})}{1 - \Pr(\vartheta = \theta_{H_0})} \underline{B}\right)^{-1}\right)^{-1}; \tag{2}$$

$$\underline{B} = -e\, p \ln p, \tag{3}$$

where $\underline{B}$ denotes the corresponding lower bound on B.

$\underline{\text{LFDR}}$ is relevant not only for Bayesian inference but also for frequentist inference, given an estimate of the value of the frequency-type prior probability $\Pr(\vartheta = \theta_{H_0})$ can be estimated. Such an estimate of $\Pr(\vartheta = \theta_{H_0})$ may be obtained using the results of multiple hypothesis tests in a large-scale data set or, when the data set does not have enough hypotheses for that, via meta-analysis. For example, the estimate 10/11 is based on meta-analyses relevant to psychology experiments (Benjamin et al., 2017). In that case, a frequentist may summarize the result of a statistical test by reporting an estimate of $\underline{\text{LFDR}}$ rather than p in order to use the information in the estimate of $\Pr(\vartheta = \theta_{H_0})$. That differs from estimating $\Pr(\vartheta = \theta_{H_0})$ to be the 50% default, which yields $\underline{\text{LFDR}} \approx \underline{B} = -e\, p \ln p$ when p is sufficiently small. ▲

But what if $\Pr(\vartheta = \theta_{H_0}) = 0$, as some maintain (e.g., Bernardo, 2011; McShane et al., 2019)? Then $\underline{\text{LFDR}} = 0$ regardless of the data, in which case reporting p would be much more informative, as when deciding whether there is enough evidence to conclude that $\vartheta > \theta_{H_0}$ or that $\vartheta < \theta_{H_0}$ (Cox, 1977; Bickel, 2011). That highlights a gap in frequentism: there is no continuity between reporting an estimate of LFDR in certain extreme cases and reporting p in other extreme cases. What should be reported in less extreme cases such as those of the $\Pr(\vartheta = \theta_{H_0}) = 1/100$ or $\Pr(\vartheta = \theta_{H_0}) = 1/10$ suggested by Hurlbert and Lombardi (2009)?

One answer is the $\underline{\text{LFDR}}$-calibrated two-sided p value equal to $p_{\underline{\text{LFDR}}} = (1 - \underline{\text{LFDR}})\, p + 2\underline{\text{LFDR}}$ according to a simple method proposed in this paper. In the extreme of a very low $\underline{\text{LFDR}}$, that calibrated two-sided p value is close to the uncalibrated p value $\left(p_{\underline{\text{LFDR}}} \approx p\right)$. At the opposite extreme, $\underline{\text{LFDR}}$ is much larger than p, resulting in a calibrated two-sided p value that is about twice as large as $\underline{\text{LFDR}}$ $\left(p_{\underline{\text{LFDR}}} \approx 2\underline{\text{LFDR}}\right)$. Figure 1 displays $p_{\underline{\text{LFDR}}}$ for various values of p and $\Pr(\vartheta = \theta_{H_0})$, revealing little need to calibrate when $\Pr(\vartheta = \theta_{H_0}) = 1/100$ but much more when $\Pr(\vartheta = \theta_{H_0}) \geq 1/2$.

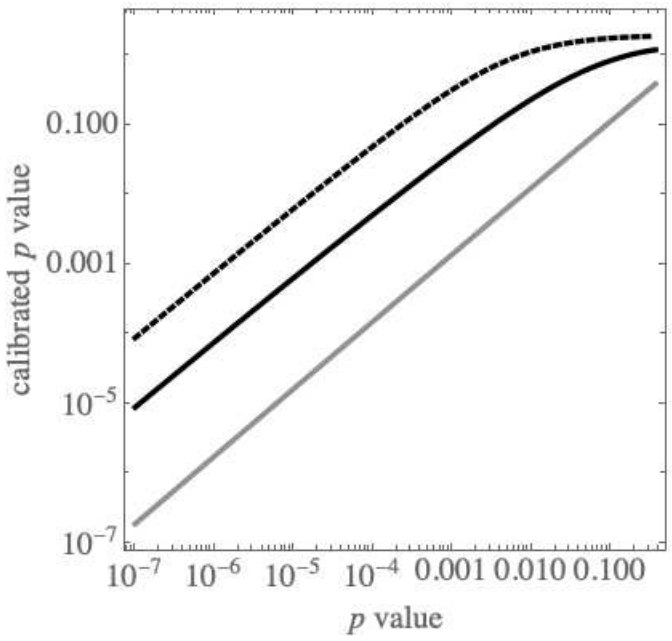

Figure 1: The LFDR-calibrated two-sided p value as a function of the uncalibrated p value. The three curves use $\Pr(\vartheta = \theta_{H_0}) = 10/11$ (dashed black), $\Pr(\vartheta = \theta_{H_0}) = 1/2$ (solid black), and $\Pr(\vartheta = \theta_{H_0}) = 1/100$ (solid gray).

That formula for p_{LFDR} emerges naturally under the empirical Bayes interpretation of confidence levels offered in Section 2. Under that interpretation, each confidence level such as 95% estimates the frequency-type posterior probability that the parameter lies within the symmetric 95% calibrated confidence interval. That estimate only applies directly when $\Pr(\vartheta = \theta_{H_0}) = 0$ but also applies indirectly conditional on $\vartheta \neq \theta_{H_0}$, yielding symmetric and one-sided confidence intervals that are calibrated by LFDR or by another estimate of LFDR. The symmetric 95% calibrated confidence intervals tend to be much shorter than the corresponding 99.5% confidence intervals (Pace and Salvan, 2019) inspired by the proposal of Benjamin et al. (2017) to test null hypotheses at the 0.5% level.

Rejecting the null hypothesis whenever θ_{H_0} is not in a symmetric 95% calibrated confidence interval based on LFDR is equivalent to rejecting the null hypothesis whenever p_{LFDR} is less than 0.05. A general statement of that result for estimates of LFDR other than LFDR and for levels of confidence other than 95% is derived in Section 3. Its example estimates of LFDR include LFDR, another lower bound on LFDR, and estimates that are not lower bounds.

Section 4 derives a point estimate of the parameter of interest from the symmetric 0% calibrated confidence interval. Technical details on confidence levels as estimates of posterior probability and on the empirical Bayes coverage of calibrated confidence intervals appear in Appendix A and in

Appendix B, respectively.

2 Calibrating confidence intervals with empirical Bayes estimation

Sections 2.1-2.3 define the preliminary concepts for Section 2.4's empirical Bayes confidence intervals in the absence of multiple comparisons.

2.1 Uncalibrated confidence intervals and p values

A typical $(100\%)\,\gamma$ confidence interval $\mathcal{C}(\gamma; Y)$ is *valid* in the sense that

$$\Pr(\mathcal{C}(\gamma; Y) \ni \theta | \vartheta = \theta) \geq \gamma, \tag{4}$$

where Y is a random sample of data distinguished from the observed sample y, and is *nested* in the sense that every confidence interval of level γ strictly contains every confidence interval of a confidence level less than γ.

Since the probability distribution Pr is unknown, its probability statements about parameter values must be estimated. For that, let the estimated posterior probability that $\vartheta \in \mathcal{C}(\gamma; y)$ be the confidence level of $\underline{\theta}(\gamma; Y) \leq \theta \leq \overline{\theta}(\gamma; Y)$, that is,

$$\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \,|\, P = p) = \gamma. \tag{5}$$

Although the fiducial argument for equation (5) when the estimand $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \,|\, P = p)$ is a subjective probability has been discredited, equation (5) has support from the broadly applicable approximation of one-sided p values to posterior probabilities (Appendix A). If $\mathcal{C}(\gamma; Y)$ is *exact* in the sense that formula (4) holds with equality, then the function $\widehat{\Pr}$ is a probability distribution known as a confidence distribution (Schweder and Hjort, 2002; Singh et al., 2005; Nadarajah et al., 2015; Schweder and Hjort, 2016; Bickel, 2019b).

Examples in which equation (5) is practical include methods of $(100\%)\,\gamma$ nested and valid confidence intervals of each of these forms:

$$\mathcal{C}_\tau(\gamma; y) = [\tau(\gamma), \infty[= \{\theta_{H_0} : p_>(\theta_{H_0}) \geq 1 - \gamma\} \tag{6}$$

$$\mathcal{C}_{\upsilon}(\gamma; y) =]-\infty, \upsilon(\gamma)] = \{\theta_{H_0} : p_{<}(\theta_{H_0}) \geq 1-\gamma\} \tag{7}$$

where $\tau(\gamma) < \upsilon(\gamma)$, $p_{>}(\theta_{H_0})$ is a p value testing the null hypothesis that $\theta = \theta_{H_0}$ with $\theta > \theta_{H_0}$ as the alternative hypothesis, and $p_{<}(\theta_{H_0})$ is a p value testing the null hypothesis that $\theta = \theta_{H_0}$ with $\theta < \theta_{H_0}$ as the alternative hypothesis.

The resulting $p^{\neq}(\theta_{H_0}) = 2\min(p_{>}(\theta_{H_0}), p_{<}(\theta_{H_0}))$ is a two-sided p value testing the null hypothesis that $\theta = \theta_{H_0}$ with $\theta \neq \theta_{H_0}$ as the alternative hypothesis. By equations (6) and (7), if $\gamma = 1 - \alpha/2$, then

$$\begin{aligned}
[\tau(\gamma), \upsilon(\gamma)] &= \{\theta_{H_0} : \tau(\gamma) \leq \theta_{H_0} \leq \upsilon(\gamma)\} \\
&= \mathcal{C}_{\tau}(\gamma; y) \cap \mathcal{C}_{\upsilon}(\gamma; y) \\
&= \{\theta_{H_0} : p_{>}(\theta_{H_0}) \geq 1-\gamma, p_{<}(\theta_{H_0}) \geq 1-\gamma\}.
\end{aligned}$$

$$\begin{aligned}
\therefore \theta_{H_0} \notin [\tau(\gamma), \upsilon(\gamma)] &\iff \theta_{H_0} \notin \{\theta_{H_0} : p_{>}(\theta_{H_0}) \geq 1-\gamma, p_{<}(\theta_{H_0}) \geq 1-\gamma\} \\
&\iff \theta_{H_0} \in \{\theta_{H_0} : p_{>}(\theta_{H_0}) < 1-\gamma\} \cup \{\theta_{H_0} : p_{<}(\theta_{H_0}) < 1-\gamma\} \\
&\iff \min(p_{>}(\theta_{H_0}), p_{<}(\theta_{H_0})) < 1-\gamma \\
&\iff p^{\neq}(\theta_{H_0}) < 2(1-\gamma) = \alpha,
\end{aligned} \tag{8}$$

which implies that $\theta_{H_0} \in [\tau(1-\alpha/2), \upsilon(1-\alpha/2)]$ is equivalent to $p^{\neq}(\theta_{H_0}) \geq \alpha$.

The confidence intervals $[\tau(\gamma), \infty[$, $]-\infty, \upsilon(\gamma)]$, and $[\tau(\gamma), \upsilon(\gamma)]$ are nested since $\tau(\gamma)$ and $\upsilon(\gamma)$ are strictly monotonic as functions of γ for every γ between 0 and 1, with $\tau(\gamma)$ decreasing and $\upsilon(\gamma)$ increasing with γ. Those confidence intervals are also valid since the probability that each covers the fixed true value of the parameter over repeated sampling is at least γ.

2.2 Estimating the probability of coverage in the presence of a plausible null hypothesis

Suppose the null hypothesis that $\vartheta = \theta_{H_0}$ is plausible enough that its frequency-type prior probability is not 0: $\Pr(\vartheta = \theta_{H_0}) > 0$. Then its LFDR, defined by $\text{LFDR} = \Pr(\vartheta = \theta_{H_0} \mid P = p)$, is also nonzero: $\text{LFDR} > 0$. That can strongly conflict with the confidence-based estimation of Section 2.1, as can be seen by considering a confidence interval $\mathcal{C}(\gamma; y)$ that covers the null hypothesis value θ_{H_0} (that is, $\theta_{H_0} \in \mathcal{C}(\gamma; y)$) and that has a level of confidence much lower than the LFDR (that is, $\gamma \ll \text{LFDR}$). Then the default confidence-based estimate of $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid y)$ is

only $\widehat{\Pr}(\mathcal{C}(\gamma; y) \mid y) = \gamma$, but $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p)$ must be at least $\Pr(\vartheta = \theta_{H_0} \mid P = p)$ since $\theta_{H_0} \in \mathcal{C}(\gamma; y)$, with the result that γ is inaccurate as an estimator of $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p)$:

$$\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid y) = \gamma \ll \text{LFDR} = \Pr(\vartheta = \theta_{H_0} \mid P = p) \leq \Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p). \tag{9}$$

For that reason, the default confidence-based estimate should be reserved for cases in which the null hypothesis has 0 frequency-type prior probability (Appendix A). That suggests considering γ as an estimate of $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid \vartheta \neq \theta_{H_0}, P = p)$, which is the posterior probability of coverage conditional on the alternative hypothesis ($\vartheta \neq \theta_{H_0}$) and on the observed p value.

Accordingly, $\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid \vartheta \neq \theta_{H_0}, P = p) = \gamma$ will denote the default confidence-based estimate of $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid \vartheta \neq \theta_{H_0}, P = p)$. $\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p)$ will instead denote an estimate of

$$\begin{aligned}
\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p) &= \Pr(\vartheta = \theta_{H_0} \mid P = p) \Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid \vartheta = \theta_{H_0}, P = p) \\
&\quad + \Pr(\vartheta \neq \theta_{H_0} \mid P = p) \Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid \vartheta \neq \theta_{H_0}, P = p) \\
&= \text{LFDR}\, \chi(\theta_{H_0} \in \mathcal{C}(\gamma; y)) + (1 - \text{LFDR}) \Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid \vartheta \neq \theta_{H_0}, P = p),
\end{aligned}$$

where $\chi(\theta_{H_0} \in \mathcal{C}(\gamma; y))$ is 1 if $\theta_{H_0} \in \mathcal{C}(\gamma; y)$ but is 0 if not. Then the natural estimate is

$$\begin{aligned}
\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid p) &= \widehat{\text{LFDR}}\, \chi(\theta_{H_0} \in \mathcal{C}(\gamma; y)) + \left(1 - \widehat{\text{LFDR}}\right) \widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid \vartheta \neq \theta_{H_0}, P = p) \\
&= \widehat{\text{LFDR}}\, \chi(\theta_{H_0} \in \mathcal{C}(\gamma; y)) + \left(1 - \widehat{\text{LFDR}}\right) \gamma,
\end{aligned} \tag{10}$$

where $\widehat{\text{LFDR}}$ is an estimate of LFDR such as one of those in Section 2.3.

2.3 Estimating the local false discovery rate

Recall that $\widehat{\text{LFDR}}$ denotes any estimate of the local false discovery rate, the posterior probability defined in equation (1). While Bickel (2012b) used results of multiple tests to determine $\widehat{\text{LFDR}}$ for each test, as is usual in empirical Bayes testing (Efron, 2010; Bickel, 2019c), the following examples of $\widehat{\text{LFDR}}$ apply when testing a single null hypothesis given $\widehat{\Pr}(\vartheta = \theta_{H_0})$, an estimate of the frequency-type probability probability that the null hypothesis holds. Those versions of $\widehat{\text{LFDR}}$ are compared in Figure 2.

Example 2. Suppose $\widehat{\text{LFDR}} = \underline{\text{LFDR}}$, where $\underline{\text{LFDR}}$ is Example 1's lower bound on LFDR. One reason for that choice of an estimate of LFDR is that the lower bound brings the estimate as

close as possible to 0 (Bickel, 2019d), which is the value of the LFDR under a null hypothesis known with certainty to be false, as some argue is the case for all point null hypotheses involving continuous parameters (e.g., Bernardo, 2011; McShane et al., 2019). While the extreme form of that position would be hard to sustain for many applications in fields as diverse as biomedicine, genomics, genetics, particle physics, and psychology, a more moderate form is plausible. In particular, Hurlbert and Lombardi (2009) argue that $\Pr(\vartheta = \theta_{H_0})$ is often much closer to 0 than to 1/2. In short, setting $\widehat{\text{LFDR}} = \underline{\text{LFDR}}$ might be justified as a way to compensate for a positive bias in $\widehat{\Pr}(\vartheta = \theta_{H_0})$.

Another reason for setting $\widehat{\text{LFDR}} = \underline{\text{LFDR}}$ is that $\underline{\text{LFDR}}$ is the maximum-likelihood estimate (MLE) of LFDR under typical conditions (Held and Ott, 2018; cf. Bickel, 2014, 2019g; Bickel, 2019c, chapter 7). Then the problem would be reduced to justifying the use of the MLE on the basis of a single observed p value: since the effective sample size is 1, the usual asymptotic justifications do not apply. Bickel (2017) used an MLE of LFDR to generate a Bayes-frequentist continuum between the extremes of credible intervals and confidence intervals.

Further arguments for $\widehat{\text{LFDR}} = \underline{\text{LFDR}}$ appear in Bickel (2019e). ▲

Example 3. Different assumptions lead to different versions of $\underline{B}$, the lower bound on the Bayes factor (Held and Ott, 2018). Instead of the version given by equation (3), this example uses $\underline{B} = e^{-z^2}$, which Held and Ott (2016) call the *universal lower bound*, where z is the standard normal quantile of a one-sided p value testing $\vartheta = \theta_{H_0}$ (Bickel, 2019g). Since that $\underline{B}$ may be too low to be a reasonable estimate of the Bayes factor B, it might require some kind of averaging with $\overline{B}$, an upper bound on B. A readily available upper bound is 1 if the two-sided p value is small enough that it can be assumed that the null hypothesis is not supported by the data. Given that the Bayes factor is between the bounds $\underline{B}$ and $\overline{B}$, the weighted geometric mean $\underline{B}^{1-c}\overline{B}^{c}$ is a minimax-optimal Bayes factor, where c is a degree of caution that is between 0 and 1 (Bickel, 2019f). For this example, let $c = 1/2$, and let the resulting *inferential Bayes factor* $\sqrt{\underline{B}\overline{B}}$ (Bickel, 2019f) serve as $\widehat{B}$, the estimate of the Bayes factor. By $\underline{B} = e^{-z^2}$ and $\overline{B} = 1$, we have $\widehat{B} = \sqrt{e^{-z^2} \times 1} = e^{-|z|}$ for a sufficiently small two-sided p value, that is, for a sufficiently large value of $|z|$. Then plugging $\widehat{B}$ into B and $\widehat{\Pr}(\vartheta = \theta_{H_0})$ into $\Pr(\vartheta = \theta_{H_0})$ in equation (1) results in the *inferential LFDR* as $\widehat{\text{LFDR}}$, the estimate of the LFDR. ▲

Example 4. A lower bound similar to that of equation (3) is the normal-distribution bound $\underline{B} = |z|\, e^{-\frac{z^2-1}{2}}$ (Held and Ott, 2016), where z, assumed to satisfy $|z| > 1$, is defined in Example 3. An interpretation of Occam's razor (Bickel, 2019a) leads to increasing $\underline{B}$ by a factor of $|z|$ (Bickel, 2019g). Then concerns about using a lower bound on B as an estimate of B may motivate instead

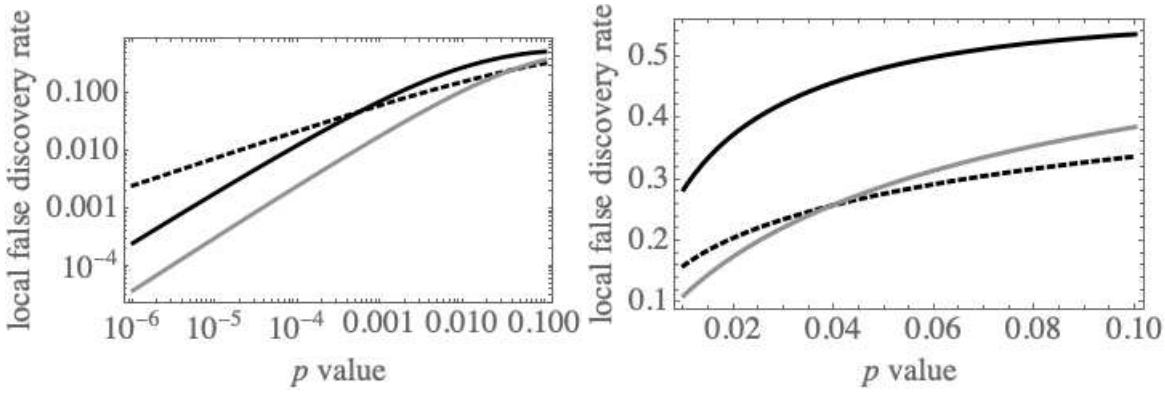

Figure 2: Each local false discovery rate estimate $\widehat{\text{LFDR}}$ as a function of the two-sided p value. The three curves are the conventional lower bound on the LFDR (solid gray), the inferential LFDR (dashed black), and the razor-based LFDR (solid black) as defined in Examples 2, 3, and 4, respectively. The left and right plots are identical except that the left plot is logarithmic, and the right plot is linear. The frequency-type prior probability estimate of $\widehat{\Pr}(\vartheta = \theta_{H_0}) = 1/2$ is assumed here.

using $\widehat{B} = |z|\,\underline{B}$ as the estimate of the Bayes factor. Substitutions of that and $\widehat{\Pr}(\vartheta = \theta_{H_0})$ in equation (1) generate the *razor-based LFDR* as $\widehat{\text{LFDR}}$. ▲

2.4 Symmetric calibrated confidence intervals

In the framework developed in Sections 2.1-2.3, the *symmetric* $100\,c\,\%$ *confidence interval* is $\left[\underline{\theta}(c;y), \overline{\theta}(c;y)\right]$, where

$$\underline{\theta}(c;y) = \tau\left(1 - \frac{1-c}{2}\right) = \tau\left(\frac{1+c}{2}\right) \tag{11}$$

$$\overline{\theta}(c;y) = \upsilon\left(1 - \frac{1-c}{2}\right) = \upsilon\left(\frac{1+c}{2}\right). \tag{12}$$

Similarly, the *symmetric* $100\,c\,\%$ $\widehat{\text{LFDR}}$*-calibrated confidence interval* is $\left[\underline{\theta}_{\widehat{\text{LFDR}}}(c;y), \overline{\theta}_{\widehat{\text{LFDR}}}(c;y)\right]$, where

$$\underline{\theta}_{\widehat{\text{LFDR}}}(c;y) = \begin{cases} \tau(\gamma^-) & \text{if } \; c \geq 2\widehat{\text{LFDR}} - 1, \tau(\gamma^-) < \theta_{H_0} \\ \tau(\gamma^+) & \text{if } \; c \leq 1 - 2\widehat{\text{LFDR}}, \tau(\gamma^+) > \theta_{H_0} \\ \theta_{H_0} & \text{otherwise} \end{cases} \tag{13}$$

$$\overline{\theta}_{\widehat{\text{LFDR}}}(c;y) = \begin{cases} \upsilon(\gamma^+) & \text{if if } \; c \leq 1 - 2\widehat{\text{LFDR}}, \upsilon(\gamma^+) < \theta_{H_0} \\ \upsilon(\gamma^-) & \text{if } \; c \geq 2\widehat{\text{LFDR}} - 1, \upsilon(\gamma^-) > \theta_{H_0} \\ \theta_{H_0} & \text{otherwise} \end{cases} \tag{14}$$

$$\gamma^- = \frac{(1+c)/2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}$$
$$\gamma^+ = \frac{(1+c)/2}{1 - \widehat{\text{LFDR}}}.$$

The equations are ready for use since $\tau(\gamma^-)$, $\tau(\gamma^+)$, $\upsilon(\gamma^-)$, and $\upsilon(\gamma^+)$ are generally available in software given that $0 \leq \gamma^- \leq 1$ and $0 \leq \gamma^+ \leq 1$, which are respectively satisfied under the $c \geq 2\widehat{\text{LFDR}} - 1$ and $c \leq 1 - 2\widehat{\text{LFDR}}$ conditions of equations (13)-(14).

Example 5. To get an 80% $\underline{\text{LFDR}}$-calibrated confidence interval of a real-valued parameter that is 0 under the null hypothesis ($\theta_{H_0} = 0$), first compute $\underline{\text{LFDR}}$ from equations (3) and (2) as per Example 2. Suppose $\underline{\text{LFDR}} = 40\%$. Then $80\% = c \geq 2\widehat{\text{LFDR}} - 1 = -20\%$ and $\gamma^- = 5/6 = 83\%$, and equations (13)-(14) yield $[\min(\tau(83\%), \theta_{H_0}), \max(\upsilon(83\%), \theta_{H_0})]$ as the 80% $\underline{\text{LFDR}}$-calibrated confidence interval, where $\tau(83\%)$ and $\upsilon(83\%)$ may be obtained from software as limits of two one-sided 83% confidence intervals. ▲

If the estimated prior probability of the null hypothesis is 0, the estimated posterior probability is also 0. In that degenerate case, $\gamma^- = \gamma^+ = (1+c)/2$ since $\widehat{\text{LFDR}} = 0$, and $\left[\underline{\theta}(c; y), \overline{\theta}(c; y)\right]$ then reduces to $[\tau(\gamma^-), \upsilon(\gamma^-)]$, the symmetric $100\,c\,\%$ confidence interval of Section 2.1.

For any c between 0 and 1, an interval $\mathcal{I}$ is a $100\,c\,\%$ *empirical Bayes interval* if

$$\widehat{\Pr}(\vartheta \in \mathcal{I} | P = p) \geq c\,. \tag{15}$$

The symmetric $100\,c\,\%$ $\widehat{\text{LFDR}}$-calibrated confidence intervals are $100\,c\,\%$ empirical Bayes intervals, as seen in Corollary 1 of Appendix B.

3 Calibrating p values with empirical Bayes estimation

Recall that $p^{\neq}(\theta_{H_0})$ denotes a two-sided p value. Let $p^{\neq}_{\widehat{\text{LFDR}}} = \left(1 - \widehat{\text{LFDR}}\right) p^{\neq}(\theta_{H_0}) + 2\widehat{\text{LFDR}}$, as illustrated in Section 1 for the $\widehat{\text{LFDR}} = \underline{\text{LFDR}}$ case.

Theorem 1. *For any $p^{\neq}(\theta_{H_0})$ satisfying the conditions of Section 2.1, $p^{\neq}_{\widehat{\text{LFDR}}} \geq \alpha$ if and only if $\theta_{H_0} \in \left[\underline{\theta}_{\widehat{\text{LFDR}}}(1-\alpha; y), \overline{\theta}_{\widehat{\text{LFDR}}}(1-\alpha; y)\right]$, the symmetric $(100\%)(1-\alpha)$ $\widehat{\text{LFDR}}$-calibrated confidence interval.*

Proof. By equations (13) and (14), $\theta_{H_0} \notin \left[\underline{\theta}_{\widehat{\text{LFDR}}}(1-\alpha; y), \overline{\theta}_{\widehat{\text{LFDR}}}(1-\alpha; y)\right]$ if and only if both

$1-\alpha \leq 1-2\widehat{\text{LFDR}}$ (i.e., $2\widehat{\text{LFDR}} \leq \alpha$) and

$$\upsilon\left(\gamma^{+}\right) < \theta_{H_0} \text{ or } \tau\left(\gamma^{+}\right) > \theta_{H_0}; \tag{16}$$

$$\gamma^{+} = \frac{1-\alpha/2}{1-\widehat{\text{LFDR}}}.$$

Disjunction (16) is equivalent to this statement according to equations (6) and (7):

$$\begin{aligned}
\theta_{H_0} &\in \left]-\infty, \upsilon\left(\gamma^{+}\right)\right[\cup \left]\tau\left(\gamma^{+}\right), \infty\right[\\
&= \left\{\theta : p_{>}\left(\theta\right) < 1-\gamma^{+}\right\} \cup \left\{\theta : p_{<}\left(\theta\right) < 1-\gamma^{+}\right\} \\
&= \left\{\theta : \min\left(p_{>}\left(\theta_{H_0}\right), p_{<}\left(\theta_{H_0}\right)\right) < 1-\gamma^{+}\right\} \\
&= \left\{\theta : p^{\neq}\left(\theta\right) < 2\left(1-\gamma^{+}\right)\right\},
\end{aligned}$$

which holds if and only if

$$p^{\neq}\left(\theta_{H_0}\right) < 2\left(1-\gamma^{+}\right) = 2\left(1-\frac{1-\alpha/2}{1-\widehat{\text{LFDR}}}\right) = \frac{\alpha-2\widehat{\text{LFDR}}}{1-\widehat{\text{LFDR}}}$$

$$\left(1-\widehat{\text{LFDR}}\right) p^{\neq}\left(\theta_{H_0}\right) + 2\widehat{\text{LFDR}} < \alpha$$

$$p^{\neq}_{\widehat{\text{LFDR}}} < \alpha.$$

Assembling the above statements of equivalence, we have

$$\begin{aligned}
\theta_{H_0} \notin \left[\underline{\theta}_{\widehat{\text{LFDR}}}\left(1-\alpha; y\right), \overline{\theta}_{\widehat{\text{LFDR}}}\left(1-\alpha; y\right)\right] &\iff 2\widehat{\text{LFDR}} \leq \alpha \text{ and } p^{\neq}_{\widehat{\text{LFDR}}} < \alpha \\
&\iff p^{\neq}_{\widehat{\text{LFDR}}} < \alpha
\end{aligned}$$

since $p^{\neq}_{\widehat{\text{LFDR}}} < \alpha \implies 2\widehat{\text{LFDR}} \leq \alpha$. In other words,

$$\theta_{H_0} \in \left[\underline{\theta}_{\widehat{\text{LFDR}}}\left(1-\alpha; y\right), \overline{\theta}_{\widehat{\text{LFDR}}}\left(1-\alpha; y\right)\right] \iff p^{\neq}_{\widehat{\text{LFDR}}} \geq \alpha.$$

□

Theorem 1 justifies calling $p^{\neq}_{\widehat{\text{LFDR}}}$ the $\widehat{\text{LFDR}}$-*calibrated two-sided p value*. It may be interpreted by noting that $p^{\neq}_{\widehat{\text{LFDR}}}/2$ is an estimate of the posterior probability of a sign error or directional error (Bickel, 2019e). In addition, $p^{\neq}_{\widehat{\text{LFDR}}}$ estimates a special case of an extended evidence value (Bickel, 2019b), which is a generalization of the evidence value proposed by Pereira and Stern

(1999). Related extended evidence values are the likelihood-ratio posterior probability defined by Aitkin (2010, p. 42), the strength of evidence defined by Evans (2015, p. 114), and the two-sided posterior probability defined by Shi and Yin (2020).

4 Calibrating point estimates with empirical Bayes estimation

A simple $\widehat{\text{LFDR}}$-based point estimate of θ is the estimated posterior mean given by

$$\begin{aligned}\widehat{\theta}\left(\widehat{\text{LFDR}}\right) &= \left(1 - \frac{p^{\neq}_{\widehat{\text{LFDR}}}}{2}\right)\widehat{\theta} + \left(\widehat{\text{LFDR}}\right)\theta_{H_0} + \left(\frac{p^{\neq}_{\widehat{\text{LFDR}}}}{2} - \widehat{\text{LFDR}}\right)\theta_{H_0} \\ &= \left(1 - \frac{p^{\neq}_{\widehat{\text{LFDR}}}}{2}\right)\widehat{\theta} + \frac{p^{\neq}_{\widehat{\text{LFDR}}}}{2}\,\theta_{H_0},\end{aligned}$$

where $\widehat{\theta}$ is the maximum-likelihood estimate or another estimate of the posterior mean conditional on the alternative hypothesis (Bickel, 2019e). However, $\widehat{\theta}\left(\widehat{\text{LFDR}}\right)$ is almost never equal to θ_{H_0}, the value of the parameter under the null hypothesis, even when $\widehat{\text{LFDR}}$, the frequency-type posterior probability of the null hypothesis, is very high. To overcome that drawback, an alternative estimate will be derived from the $\widehat{\text{LFDR}}$-calibrated confidence intervals of Section 2.4.

The main idea is to define a point estimate that would be in the symmetric $100\,c\,\%$ $\widehat{\text{LFDR}}$-calibrated confidence interval for every value of c, in agreement with the argument for 0% confidence intervals reviewed by Pace and Salvan (1997, Appendix D). Since the intervals are nested, that requirement is equivalent to the requirement that the point estimate be in the symmetric 0% $\widehat{\text{LFDR}}$-calibrated confidence interval, which is $\left[\underline{\theta}_{\widehat{\text{LFDR}}}(0; y), \overline{\theta}_{\widehat{\text{LFDR}}}(0; y)\right]$, where

$$\underline{\theta}_{\widehat{\text{LFDR}}}(0; y) = \tau_{\widehat{\text{LFDR}}}(1/2) = \begin{cases} \tau\left(\frac{1/2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \widehat{\text{LFDR}} \leq \frac{1}{2}, \tau\left(\frac{1/2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\ \tau\left(\frac{1/2}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \widehat{\text{LFDR}} \leq \frac{1}{2}, \tau\left(\frac{1/2}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\ \theta_{H_0} & \text{otherwise} \end{cases}$$

$$\overline{\theta}_{\widehat{\text{LFDR}}}(0; y) = \upsilon_{\widehat{\text{LFDR}}}(1/2) = \begin{cases} \upsilon\left(\frac{1/2}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \widehat{\text{LFDR}} \leq \frac{1}{2}, \upsilon\left(\frac{1/2}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\ \upsilon\left(\frac{1/2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \widehat{\text{LFDR}} \leq \frac{1}{2}, \upsilon\left(\frac{1/2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\ \theta_{H_0} & \text{otherwise} \end{cases}$$

according to equations (13) and (14) with $c = 0$.

It can be seen that if $\widehat{\Pr}(\vartheta = \theta_{H_0}) > 0$, then $\widehat{\text{LFDR}} > 0$, and $\underline{\theta}_{\widehat{\text{LFDR}}}(0; y) = \overline{\theta}_{\widehat{\text{LFDR}}}(0; y)$ could only occur if $\theta_{H_0} = \underline{\theta}(0; y) = \overline{\theta}(0; y)$, which does not happen since θ_{H_0} does not depend on y. By contrast, in the degenerate case that $\widehat{\Pr}(\vartheta = \theta_{H_0}) = 0$, we have $\widehat{\text{LFDR}} = 0$, and $\left[\underline{\theta}_{\widehat{\text{LFDR}}}(0; y), \overline{\theta}_{\widehat{\text{LFDR}}}(0; y)\right]$ is the confidence interval $\left[\underline{\theta}(0; y), \overline{\theta}(0; y)\right]$, often enabling $\underline{\theta}_{\widehat{\text{LFDR}}}(0; y) = \overline{\theta}_{\widehat{\text{LFDR}}}(0; y)$ since the limits of typical 0% confidence intervals from continuous data are equal to each other (e.g., Pace and Salvan, 1997, Appendix D). In cases of $\widehat{\Pr}(\vartheta = \theta_{H_0}) > 0$, an additional condition is needed to uniquely define a point estimate. To exercise caution in case the null hypothesis holds, the $\widehat{\text{LFDR}}$-*calibrated point estimate* is the value in $\left[\underline{\theta}_{\widehat{\text{LFDR}}}(0; y), \overline{\theta}_{\widehat{\text{LFDR}}}(0; y)\right]$ that is closest to θ_{H_0}:

$$\widehat{\theta}_{\widehat{\text{LFDR}}} = \begin{cases} \upsilon\left(\frac{1/2}{1-\widehat{\text{LFDR}}}\right) & \text{if } \widehat{\text{LFDR}} \leq \frac{1}{2}, \upsilon\left(\frac{1/2}{1-\widehat{\text{LFDR}}}\right) < \theta_{H_0} \\ \tau\left(\frac{1/2}{1-\widehat{\text{LFDR}}}\right) & \text{if } \widehat{\text{LFDR}} \leq \frac{1}{2}, \tau\left(\frac{1/2}{1-\widehat{\text{LFDR}}}\right) > \theta_{H_0} \\ \theta_{H_0} & \text{otherwise} \end{cases}.$$

A less cautious but also less biased approach instead takes the midpoint of $\underline{\theta}_{\widehat{\text{LFDR}}}(0; y)$ and $\overline{\theta}_{\widehat{\text{LFDR}}}(0; y)$ as the point estimate. The $\widehat{\text{LFDR}}$-calibrated point estimate resembles the posterior median that Bickel (2012b) applied to genomics data using a version of $\widehat{\text{LFDR}}$ developed for multiple testing.

Acknowledgments

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

A Confidence levels as estimates of frequency-type probabilities

Confidence intervals are often interpreted as interval-valued estimates of parameter values. The idea has been extended to entire confidence distributions as distribution-valued estimates of parameter values (Singh et al., 2007; Xie and Singh, 2013). But what does the confidence level of a confidence interval estimate? For example, given an observed 95% confidence interval $[\theta(2.5\%), \theta(97.5\%)]$, what is the number 95% an estimate for? One answer is that it estimates the indicator of whether θ, the true value of the parameter, is in the confidence interval, where the value of the indicator is

1 if $\theta \in [\theta(2.5\%), \theta(97.5\%)]$ but is 0 if $\theta \notin [\theta(2.5\%), \theta(97.5\%)]$ (Bickel, 2012a). While that may work for a fixed parameter value, it needs to be generalized to cases in which a random parameter ϑ has an unknown frequency-type distribution.

How can the confidence level help us estimate the the frequency-type posterior probability that the parameter lies within the observed confidence interval? If very little is known about that distribution, it may be reasonable to estimate the posterior probability with an estimator that is higher for higher confidence levels. That is, the estimated posterior probability that the parameter is in the observed confidence interval increases when, for example, the confidence level is increased from 95% to 99% since that results in a wider interval. For example, an estimator $\widehat{\Pr}([\tau(\gamma), \upsilon(\gamma)] \mid P = p)$ of the frequency-type posterior probability $\Pr([\tau(\gamma), \upsilon(\gamma)] \mid P = p)$ that increases monotonically with the confidence level γ of the observed confidence interval $[\tau(\gamma), \upsilon(\gamma)]$.

That can be made more definite using the observation that one-sided p values tend to closely approximate posterior probabilities corresponding to certain prior probability density functions that are diffuse (Pratt, 1965, §7) or impartial (Casella and Berger, 1987, §2); see Shi and Yin (2020). Accordingly, in the notation of Section 2.1, let the estimators of $\Pr(\vartheta \geq \theta_{H_0} \mid P = p)$ and $\Pr(\vartheta \leq \theta_{H_0} \mid P = p)$ be

$$\widehat{\Pr}(\vartheta \geq \theta_{H_0} \mid P = p) = p_{<}(\theta_{H_0}) \tag{17}$$

$$\widehat{\Pr}(\vartheta \leq \theta_{H_0} \mid P = p) = p_{>}(\theta_{H_0}) \tag{18}$$

for any θ_{H_0}, where the inequalities may be replaced by strict inequalities since $\widehat{\Pr}(\vartheta = \theta_{H_0} \mid P = p) = 0$. The fact that

$$\Pr\left(\underline{\theta} \leq \vartheta \leq \overline{\theta} \middle| P = p\right) = 1 - \Pr(\vartheta < \underline{\theta} \middle| P = p) - \Pr\left(\vartheta > \overline{\theta} \middle| P = p\right)$$

for any interval $\left[\underline{\theta}, \overline{\theta}\right]$ of parameter values suggests estimating that probability by

$$\widehat{\Pr}\left(\underline{\theta} \leq \vartheta \leq \overline{\theta} \middle| P = p\right) = 1 - \widehat{\Pr}(\vartheta < \underline{\theta} \middle| P = p) - \widehat{\Pr}\left(\vartheta > \overline{\theta} \middle| P = p\right). \tag{19}$$

Theorem 2. *If $p_{>}(\theta_{H_0}) = 1 - p_{<}(\theta_{H_0})$ is a bijective function of θ_{H_0}, then the observed confidence intervals of Section 2.1 satisfy*

$$\widehat{\Pr}(\vartheta \in [\tau(\gamma), \infty[\mid P = p) = \gamma \tag{20}$$

$$\widehat{\Pr}(\vartheta \in]-\infty, \upsilon(\gamma)] \mid P = p) = \gamma \tag{21}$$

$$\widehat{\Pr}\left(\vartheta \in \left[\tau\left(\frac{1+\gamma}{2}\right), \upsilon\left(\frac{1+\gamma}{2}\right)\right] \mid P = p\right) = \gamma$$

for any $\gamma \in [0,1]$.

Proof. From equations (18), (6), and $p_{>}(\theta_{H_0}) = 1 - p_{<}(\theta_{H_0})$,

$$\begin{aligned}
\widehat{\Pr}(\vartheta \in [\tau(\gamma), \infty[\mid P = p) &= p_{<}(\tau(\gamma)) \\
&= p_{<}(\inf\{\theta_{H_0} : p_{>}(\theta_{H_0}) \geq 1-\gamma\}) \\
&= p_{<}(\inf\{\theta_{H_0} : p_{<}(\theta_{H_0}) < \gamma\}) \\
&= p_{<}\left(p_{<}^{-1}(\gamma)\right) = \gamma, \qquad (22)
\end{aligned}$$

where equation (22) follows from the bijectivity condition. Analogously, from equations (17), (7), and $p_{>}(\theta_{H_0}) = 1 - p_{<}(\theta_{H_0})$,

$$\begin{aligned}
\widehat{\Pr}(\vartheta \in]-\infty, \upsilon(\gamma)] \mid P = p) &= p_{>}(\upsilon(\gamma)) \\
&= p_{>}(\sup\{\theta_{H_0} : p_{<}(\theta_{H_0}) \geq 1-\gamma\}) \\
&= p_{>}(\sup\{\theta_{H_0} : p_{>}(\theta_{H_0}) < \gamma\}) \\
&= p_{>}\left(p_{>}^{-1}(\gamma)\right) = \gamma, \qquad (23)
\end{aligned}$$

with equation (23) from bijectivity. By equation (19) and the now proved equations (20)-(21),

$$\begin{aligned}
\widehat{\Pr}\left(\vartheta \in \left[\tau\left(\frac{1+\gamma}{2}\right), \upsilon\left(\frac{1+\gamma}{2}\right)\right] \mid P = p\right) &= 1 - \widehat{\Pr}\left(\vartheta < \tau\left(\frac{1+\gamma}{2}\right) \mid P = p\right) - \widehat{\Pr}\left(\vartheta > \upsilon\left(\frac{1+\gamma}{2}\right) \mid P = p\right) \\
&= 1 - \left(1 - \widehat{\Pr}\left(\vartheta \geq \tau\left(\frac{1+\gamma}{2}\right) \mid P = p\right)\right) \\
&\quad - \left(1 - \widehat{\Pr}\left(\vartheta \leq \upsilon\left(\frac{1+\gamma}{2}\right) \mid P = p\right)\right) \\
&= 1 - \left(1 - \frac{1+\gamma}{2}\right) - \left(1 - \frac{1+\gamma}{2}\right) = 1 - 2 + 1 + \gamma = \gamma.
\end{aligned}$$

□

That justifies equation (5)'s estimating $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p)$ by γ for diffuse or impartial prior distributions. On the other hand, estimating $\Pr(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p)$ with the confidence-based estimate $\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p) = \gamma$ can be highly inaccurate when it is plausible that the null hypothesis that $\vartheta = \theta_{H_0}$ is approximately true, for in that case the prior distribution flagrantly violates the regularity conditions of Pratt (1965) and Casella and Berger (1987) since

$\Pr(\vartheta = \theta_{H_0}) > 0$; see equation (9). How to apply confidence-based estimation to that setting is the topic of Section 2.2.

It is well known from studies of fiducial probability that $\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p) = \gamma$ can break rules of probability in the sense that $\widehat{\Pr}(\vartheta \in \mathcal{C}(\gamma; y) \mid P = p)$ is not necessarily an exact posterior probability (e.g., Lindley, 1958), but minor violations need not cause concern, for estimators need not satisfy the properties of the quantities they estimate. That is why Wilkinson (1977, §6.2) considered γ as an estimated level of belief (Bickel, 2019d,e). Here, it is instead considered as an estimated frequency-type probability in the sense of Section 1 on the basis of Theorem 2.

B Calibrated confidence intervals

For any confidence levels γ_1 and γ_2 such that $1 \leq \gamma_1 + \gamma_2 \leq 2$, a generalization of the symmetric $(100\%)(\gamma_1 + \gamma_2 - 1)$ $\widehat{\text{LFDR}}$-calibrated confidence intervals of Section 2.4 is the $(100\%)(\gamma_1 + \gamma_2 - 1)$ $\widehat{\text{LFDR}}$-*calibrated confidence interval* given by $\left[\tau_{\widehat{\text{LFDR}}}(\gamma_1), \upsilon_{\widehat{\text{LFDR}}}(\gamma_2)\right]$, where

$$\tau_{\widehat{\text{LFDR}}}(\gamma_1) = \begin{cases} \tau\left(\frac{\gamma_1 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \gamma_1 \geq \widehat{\text{LFDR}}, \tau\left(\frac{\gamma_1 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\ \tau\left(\frac{\gamma_1}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \gamma_1 \leq 1 - \widehat{\text{LFDR}}, \tau\left(\frac{\gamma_1}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\ \theta_{H_0} & \text{otherwise} \end{cases} \tag{24}$$

$$\upsilon_{\widehat{\text{LFDR}}}(\gamma_2) = \begin{cases} \upsilon\left(\frac{\gamma_2}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \gamma_2 \leq 1 - \widehat{\text{LFDR}}, \upsilon\left(\frac{\gamma_2}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\ \upsilon\left(\frac{\gamma_2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) & \text{if } \gamma_2 \geq \widehat{\text{LFDR}}, \upsilon\left(\frac{\gamma_2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\ \theta_{H_0} & \text{otherwise} \end{cases} . \tag{25}$$

That is an empirical Bayes interval as defined by Section 2.4's equation (15).

Theorem 3. *For every* $\gamma_1 \in [0, 1]$ *and* $\gamma_2 \in [0, 1]$ *satisfying* $1 \leq \gamma_1 + \gamma_2 \leq 2$, $\left[\tau_{\widehat{\text{LFDR}}}(\gamma_1), \upsilon_{\widehat{\text{LFDR}}}(\gamma_2)\right]$ *is a* $(100\%)(\gamma_1 + \gamma_2 - 1)$ *empirical Bayes interval. In the special case that* $\widehat{\Pr}(\vartheta = \theta) = 0$, $\left[\tau_{\widehat{\text{LFDR}}}(\gamma_1), \upsilon_{\widehat{\text{LFDR}}}(\gamma_2)\right]$ *is the valid* $(100\%)(\gamma_1 + \gamma_2 - 1)$ *confidence interval*

$$[\tau_0(\gamma_1), \upsilon_0(\gamma_2)] = [\tau(\gamma_1), \upsilon(\gamma_2)]. \tag{26}$$

Proof. By equation (24),

$$\begin{aligned}
\widehat{\Pr}\left(\vartheta < \tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right) \mid p\right) &= \widehat{\Pr}\left(\vartheta = \theta_{H_0} \mid P = p\right) \widehat{\Pr}\left(\vartheta < \tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right) \mid \vartheta = \theta_{H_0}, p\right) + \widehat{\Pr}\left(\vartheta \neq \theta_{H_0} \mid P = p\right) \widehat{\Pr}\left(\vartheta < \tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right)\right. \\
&= \widehat{\text{LFDR}} \chi\left(\tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right) > \theta_{H_0}\right) + \left(1 - \widehat{\text{LFDR}}\right)\left(1 - \tau^{-1}\left(\tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right)\right)\right) \\
&= \begin{cases}
\widehat{\text{LFDR}} \times 0 + \left(1 - \widehat{\text{LFDR}}\right)\left(1 - \tau^{-1}\left(\tau\left(\frac{\gamma_1 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right)\right)\right) & \text{if } \gamma_1 \geq \widehat{\text{LFDR}}, \\
& \tau\left(\frac{\gamma_1 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\
\widehat{\text{LFDR}} \times 1 + \left(1 - \widehat{\text{LFDR}}\right)\left(1 - \tau^{-1}\left(\tau\left(\frac{\gamma_1}{1 - \widehat{\text{LFDR}}}\right)\right)\right) & \text{if } \gamma_1 \leq 1 - \widehat{\text{LFDR}}, \\
& \tau\left(\frac{\gamma_1}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\
\widehat{\text{LFDR}} \times 0 + \left(1 - \widehat{\text{LFDR}}\right)\left(1 - \tau^{-1}\left(\theta_{H_0}\right)\right) & \text{otherwise}
\end{cases} \\
&= \begin{cases}
1 - \gamma_1 & \text{if } \gamma_1 \geq \widehat{\text{LFDR}}, \tau\left(\frac{\gamma_1 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\
1 - \gamma_1 & \text{if } \gamma_1 \leq 1 - \widehat{\text{LFDR}}, \tau\left(\frac{\gamma_1}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\
\left(1 - \widehat{\text{LFDR}}\right)\left(1 - \tau^{-1}\left(\theta_{H_0}\right)\right) & \text{otherwise}
\end{cases},
\end{aligned}$$

which in all cases is $1 - \gamma_1$ or lower, yielding $\widehat{\Pr}\left(\vartheta < \tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right) \mid p\right) \leq 1 - \gamma_1$. Likewise, equation (25) gives

$$\begin{aligned}
\widehat{\Pr}\left(\vartheta \leq \upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right) \mid p\right) &= \widehat{\text{LFDR}} \chi\left(\upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right) \geq \theta_{H_0}\right) + \left(1 - \widehat{\text{LFDR}}\right) \upsilon^{-1}\left(\upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right)\right) \\
&= \begin{cases}
\widehat{\text{LFDR}} \times 0 + \left(1 - \widehat{\text{LFDR}}\right) \upsilon^{-1}\left(\upsilon\left(\frac{\gamma_2}{1 - \widehat{\text{LFDR}}}\right)\right) & \text{if } \gamma_2 \leq 1 - \widehat{\text{LFDR}}, \upsilon\left(\frac{\gamma_2}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\
\widehat{\text{LFDR}} \times 1 + \left(1 - \widehat{\text{LFDR}}\right) \upsilon^{-1}\left(\upsilon\left(\frac{\gamma_2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right)\right) & \text{if } \gamma_2 \geq \widehat{\text{LFDR}}, \upsilon\left(\frac{\gamma_2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\
\widehat{\text{LFDR}} \times 1 + \left(1 - \widehat{\text{LFDR}}\right) \upsilon^{-1}\left(\theta_{H_0}\right) & \text{otherwise}
\end{cases} \\
&= \begin{cases}
\gamma_2 & \text{if } \gamma_2 \leq 1 - \widehat{\text{LFDR}}, \upsilon\left(\frac{\gamma_2}{1 - \widehat{\text{LFDR}}}\right) < \theta_{H_0} \\
\gamma_2 & \text{if } \gamma_2 \geq \widehat{\text{LFDR}}, \upsilon\left(\frac{\gamma_2 - \widehat{\text{LFDR}}}{1 - \widehat{\text{LFDR}}}\right) > \theta_{H_0} \\
\widehat{\text{LFDR}} + \left(1 - \widehat{\text{LFDR}}\right) \upsilon^{-1}\left(\theta_{H_0}\right) & \text{otherwise}
\end{cases},
\end{aligned}$$

which in all cases is γ_2 or higher, yielding $\widehat{\Pr}\left(\vartheta \leq \upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right) \mid p\right) \geq \gamma_2$. The two inequalities give

$$\begin{aligned}
\widehat{\Pr}\left(\vartheta \in \left[\tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right), \upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right)\right] \mid p\right) &= 1 - \widehat{\Pr}\left(\vartheta < \tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right) \mid p\right) - \widehat{\Pr}\left(\vartheta > \upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right) \mid p\right) \\
&= 1 - \widehat{\Pr}\left(\vartheta < \tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right) \mid p\right) - \left(1 - \widehat{\Pr}\left(\vartheta \leq \upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right) \mid p\right)\right) \\
&= \widehat{\Pr}\left(\vartheta \leq \upsilon_{\widehat{\text{LFDR}}}\left(\gamma_2\right) \mid p\right) - \widehat{\Pr}\left(\vartheta < \tau_{\widehat{\text{LFDR}}}\left(\gamma_1\right) \mid p\right) \geq \gamma_2 - \left(1 - \gamma_1\right).
\end{aligned}$$

That satisfies equation (15), establishing the claim that $\left[\tau_{\widehat{\text{LFDR}}}(\gamma_1), \upsilon_{\widehat{\text{LFDR}}}(\gamma_2)\right]$ is a $(100\%)(\gamma_1+\gamma_2-1)$ empirical Bayes interval.

In the special case that $\widehat{\Pr}(\vartheta=\theta)=0$, we have $\widehat{\text{LFDR}}=0$ by Bayes's theorem, and equation (26) follows by substitution. Since $[\tau(\gamma;Y),\infty[=[\tau(\gamma),\infty[$ and $]-\infty,\upsilon(\gamma;Y)]=]-\infty,\upsilon(\gamma)]$ are valid confidence intervals,

$$\begin{aligned}\Pr\left(\tau(\gamma_1;Y)\leq\theta\leq\upsilon(\gamma_2;Y)\,|\vartheta=\theta\right)&=1-\Pr\left(\tau(\gamma_1;Y)>\theta|\vartheta=\theta\right)-\Pr\left(\upsilon(\gamma_2;Y)<\theta|\vartheta=\theta\right)\\&\geq 1-(1-\gamma_1)-(1-\gamma_2)=(\gamma_1+\gamma_2-1)\end{aligned}$$

holds for all θ, establishing the validity of $[\tau(\gamma_1),\upsilon(\gamma_2)]$ as a $(100\%)(\gamma_1+\gamma_2-1)$ confidence interval. □

The $(100\%)(\gamma_1+\gamma_2-1)$ $\widehat{\text{LFDR}}$-calibrated confidence interval $\left[\tau_{\widehat{\text{LFDR}}}(\gamma_1),\upsilon_{\widehat{\text{LFDR}}}(\gamma_2)\right]$ is essentially the $(100\%)(\gamma_1+\gamma_2-1)$ marginal confidence interval of Bickel (2012b), with the main difference being that the latter requires the choice of an empirical Bayes estimator of LFDR that is based on data across multiple comparisons.

Corollary 1. *For every $c\in[0,1]$, the symmetric $100\,c\,\%$ $\widehat{\text{LFDR}}$-calibrated confidence interval $\left[\underline{\theta}_{\widehat{\text{LFDR}}}(c;y),\overline{\theta}_{\widehat{\text{LFDR}}}(c;y)\right]$ of Section 2.4 is a $100\,c\,\%$ empirical Bayes interval.*

Proof. Substituting $\gamma_1=\gamma_2=(1+c)/2$ into equations (24)-(25) yields equations (13)-(14). The claim then follows from Theorem 3. □

References

Aitkin, M., 2010. Statistical Inference: An Integrated Bayesian/Likelihood Approach. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

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., 1997. Noninformative priors do not exist: A discussion. Journal of Statistical Planning and Inference 65, 159–189.

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

Bernardo, J. M., Smith, A. F. M., 1994. Bayesian Theory. John Wiley & Sons, New York.

Bickel, D. R., 2011. Estimating the null distribution to adjust observed confidence levels for genome-scale screening. Biometrics 67, 363–370.

Bickel, D. R., 2012a. Coherent frequentism: A decision theory based on confidence sets. Communications in Statistics - Theory and Methods 41, 1478–1496.

Bickel, D. R., 2012b. Empirical Bayes interval estimates that are conditionally equal to unadjusted confidence intervals or to default prior credibility intervals. Statistical Applications in Genetics and Molecular Biology 11 (3), art. 7.

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., 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, 10788–10799.

Bickel, D. R., 2019a. An explanatory rationale for priors sharpened into Occam's razors. Bayesian Analysis, DOI: 10.1214/19-BA1189.
URL `https://doi.org/10.1214/19-BA1189`

Bickel, D. R., 2019b. Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors. Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2019.1580739.
URL `https://doi.org/10.1080/03610926.2019.1580739`

Bickel, D. R., 2019c. Genomics Data Analysis: False Discovery Rates and Empirical Bayes Methods. Chapman and Hall/CRC, New York.
URL https://davidbickel.com/genomics/

Bickel, D. R., 2019d. Maximum entropy derived and generalized under idempotent probability to address Bayes-frequentist uncertainty and model revision uncertainty, working paper, DOI: 10.5281/zenodo.2645555.
URL https://doi.org/10.5281/zenodo.2645555

Bickel, D. R., 2019e. Null hypothesis significance testing interpreted and calibrated by estimating probabilities of sign errors: A Bayes-frequentist continuum, working paper, DOI: 10.5281/zenodo.3569888.
URL https://doi.org/10.5281/zenodo.3569888

Bickel, D. R., 2019f. Reporting Bayes factors or probabilities to decision makers of unknown loss functions. Communications in Statistics - Theory and Methods 48, 2163–2174.

Bickel, D. R., 2019g. Sharpen statistical significance: Evidence thresholds and Bayes factors sharpened into Occam's razor. Stat 8 (1), e215.

Carnap, R., 1962. Logical Foundations of Probability. University of Chicago Press, Chicago.

Carnap, R., 1971. A basic system of inductive logic, part 1. Studies in Inductive Logic and Probability, Vol. 1. University of California Press, Berkeley, pp. 3–165.

Carnap, R., 1980. A basic system of inductive logic. Vol. 2 of Studies in Inductive Logic and Probability, Vol. 2. University of California Press, Berkeley, pp. 7–155.

Casella, G., Berger, R. L., 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association 82, 106–111.

Cox, D., 1978. Foundations of statistical inference: The case for eclecticism. Journal of the Australian Statistical Society 20, 43–59.

Cox, D. R., 1977. The role of significance tests. Scandinavian Journal of Statistics 4, 49–70.

Dawid, A. P., Stone, M., Zidek, J. V., 1973. Marginalization paradoxes in Bayesian and structural inference (with discussion). Journal of the Royal Statistical Society B 35, 189–233.

Edwards, A. W. F., 1992. Likelihood. Johns Hopkins Press, Baltimore.

Efron, B., 2008. Simultaneous inference: When should hypothesis testing problems be combined? Annals of Applied Statistics 2, 197–223.

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

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151–1160.

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

Fisher, R. A., 1973. Statistical Methods and Scientific Inference. Hafner Press, New York.

Fraser, D. A. S., 2011. Is Bayes posterior just quick and dirty confidence? Statistical Science 26, 299–316.

Hacking, I., 2001. An introduction to probability and inductive logic. Cambridge University Press, Cambridge.

Hald, A., 2007. A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935. Springer, 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.

Hurlbert, S., Lombardi, C., 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46, 311–349.

Jaynes, E., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge.

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

Kass, R. E., Wasserman, L., 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association 91, 1343–1370.

Keynes, J. M., 1921. A Treatise On Probability. Cosimo Classics (2006 impression), New York.

Kyburg, H. E., Teng, C. M., 2001. Uncertain Inference. Cambridge University Press, Cambridge.

Kyburg, H. E., Teng, C. M., 2006. Nonmonotonic logic and statistical inference. Computational Intelligence 22, 26–51.

Levi, I., 2008. Degrees of belief. Journal of Logic and Computation 18, 699–719.

Lindley, D. V., 1958. Fiducial distributions and Bayes' theorem. Journal of the Royal Statistical Society B 20, 102–107.

McShane, B. B., Gal, D., Gelman, A., Robert, C., Tackett, J. L., 2019. Abandon statistical significance. The American Statistician 73 (sup1), 235–245.

Nadarajah, S., Bityukov, S., Krasnikov, N., 2015. Confidence distributions: A review. Statistical Methodology 22, 23–46.

Neyman, J., 1957. "inductive behavior" as a basic concept of philosophy of science. Revue de l'Institut International de Statistique / Review of the International Statistical Institute 25 (1/3), 7–22.

Pace, L., Salvan, A., 1997. Principles of Statistical Inference: From a Neo-Fisherian Perspective. Advanced Series on Statistical Science & Applied Probability. World Scientific, Singapore.

Pace, L., Salvan, A., 2019. Likelihood, replicability and Robbins' confidence sequences. International Statistical Review, DOI: 10.1111/insr.12355.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12355

Pereira, C. A. B., Stern, J. M., 1999. Evidence and credibility: Full Bayesian significance test for precise hypotheses. Entropy 1 (4), 99–110.

Pratt, J. W., 1965. Bayesian interpretation of standard inference statements. Journal of the Royal Statistical Society B 27, 169–203.

Robbins, H., 1956. An empirical Bayes approach to statistics. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1. University of California Press, Berkeley, pp. 157–163.

Schweder, T., Hjort, N., 2016. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Schweder, T., Hjort, N. L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29, 309–332.

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

Shi, H., Yin, G., 2020. Reconnecting p-value and posterior probability under one- and two-sided tests. The American Statistician 0 (0), 1–11, , DOI: 10.1080/00031305.2020.1717621.
URL https://doi.org/10.1080/00031305.2020.1717621

Singh, K., Xie, M., Strawderman, W. E., 2005. Combining information from independent sources through confidence distributions. Annals of Statistics 33, 159–183.

Singh, K., Xie, M., Strawderman, W. E., 2007. Confidence distribution (CD) – distribution estimator of a parameter. IMS Lecture Notes Monograph Series 2007 54, 132–150.

Wasserstein, R. L., Schirm, A. L., Lazar, N. A., 2019. Moving to a world beyond "p $<$ 0.05". The American Statistician 73 (sup1), 1–19.

Wilkinson, G. N., 1977. On resolving the controversy in statistical inference (with discussion). Journal of the Royal Statistical Society B 39, 119–171.

Williamson, J., 2010. In Defence of Objective Bayesianism. Oxford University Press, Oxford.

Xie, M.-G., Singh, K., 2013. Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review 81 (1), 3–39.

Zabell, S. L., 1992. R. A. Fisher and the fiducial argument. Statistical Science 7, 369–387.