Null hypothesis significance testing interpreted and calibrated by estimating probabilities of sign errors: A Bayes-frequentist continuum

December 10, 2019

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

Concepts from multiple testing can improve tests of single hypotheses. The proposed definition of the calibrated p value is an estimate of the local false sign rate, the posterior probability that the direction of the estimated effect is incorrect. Interpreting one-sided p values as estimates of conditional posterior probabilities, that calibrated p value is (1 - LFDR) p/2 + LFDR, where p is a two-sided p value and LFDR is an estimate of the local false discovery rate, the posterior probability that a point null hypothesis is true given p. A simple option for LFDR is the posterior probability derived from estimating the Bayes factor to be its e p ln(1/p) lower bound.

The calibration provides a continuum between significance testing and traditional Bayesian testing. The former effectively assumes the prior probability of the null hypothesis is 0, as some statisticians argue is the case. Then the calibrated p value is equal to p/2, a one-sided p value, since LFDR = 0. In traditional Bayesian testing, the prior probability of the null hypothesis is at least 50%, which usually results in LFDR >> p. At that end of the continuum, the calibrated p value is close to LFDR.

Keywords: calibrated effect size estimation; *p* value; directional error; dividing null hypothesis; replication crisis; reproducibility crisis; sign error; Type III error

1 Introduction

Meta-analyses of large numbers of previous studies from biomedicine and neuroscience have raised concerns that many published results cannot be replicated (Ioannidis, 2005; Nieuwenhuis et al., 2011; Button et al., 2013), contributing to the perceived replication crisis in many scientific fields (Begley and Ioannidis, 2015), especially psychology (Open Science Collaboration, 2015; Hughes, 2018). The statistics community has responded with guidelines on hypothesis testing and recommendations to emphasize effect sizes (e.g., Wasserstein and Lazar, 2016). However, conflicting proposals among statisticians on how to improve statistical data analysis (e.g., Wasserstein et al., 2019, and references) cause confusion among non-statisticians (Schachtman, 2019; Mayo, 2019), leaving statistical consultants with the responsibility of sifting through the arguments to provide their collaborators practical solutions.

For example, many Bayesians propose to address criticisms of null hypothesis significance testing by transforming the p value to a lower bound on the posterior probability that the null hypothesis is true: see Held and Ott (2018) and its references.

Example 1. Assuming the two-sided p value is not large ($p \leq 1/e$) when testing the null hypothesis $H_0 : \theta = \theta_{H_0}$, Sellke et al. (2001) and Benjamin and Berger (2019) recommend

$$\underline{B} = -e\, p \ln p \tag{1}$$

as a lower bound on the Bayes factor $B = \Pr(P = p \,|\theta = \theta_{H_0}) \,/\, \Pr(P = p \,|\theta \neq \theta_{H_0})$, where θ is the unknown value of the parameter of interest, θ_{H_0} is the fixed parameter value of the null hypothesis, P is the random variable representing the p value before it is observed to be equal to the number p. Since the posterior probability is

$$\Pr(\theta = \theta_{H_0} \,|\, P = p) = \frac{\Pr(\theta = \theta_{H_0}) \Pr(P = p \,|\theta = \theta_{H_0})}{\Pr(P = p)} = \left(1 + \left(\frac{\Pr(\theta = \theta_{H_0})}{1 - \Pr(\theta = \theta_{H_0})}\, B\right)^{-1}\right)^{-1} \tag{2}$$

according to Bayes's theorem, it has a lower bound of

$$v = \left(1 + \left(\frac{\Pr(\theta = \theta_{H_0})}{1 - \Pr(\theta = \theta_{H_0})}\underline{B}\right)^{-1}\right)^{-1}, \tag{3}$$

called the *v value* because a quantity approximated by $\underline{B}$ appears in Vovk (1993, §9). ▲

Since $\Pr(\theta = \theta_{H_0} \,|\, P = p)$ is typically much larger than p when $\Pr(\theta = \theta_{H_0}) \geq 1/2$, it is often

claimed that p "overstates" the strength of the evidence against the null hypothesis (e.g., Goodman, 1999). That conclusion is disputed by Hurlbert and Lombardi (2009), who argue that since prudent scientists tend to believe the null hypotheses they test are false, $\Pr(\theta = \theta_{H_0})$ should be much smaller than 1/2, perhaps 1/10 or 1/100.

In fact, Bernardo (2011), McShane et al. (2019), and others argue that since systematic errors prevent $\theta = \theta_{H_0}$ from ever being exactly true, it follows that 0 is the only reasonable value for $\Pr(\theta = \theta_{H_0})$; cf. van den Bergh et al. (2019). In that case, $\Pr(\theta = \theta_{H_0} \mid P = p) = 0$, which would make traditional Bayesian hypothesis testing useless. Frequentist hypothesis testing, on the other hand, could still serve to determine whether the sample is large enough to warrant concluding that $\theta > \theta_{H_0}$ or that $\theta < \theta_{H_0}$. In that context, $\theta = \theta_{H_0}$ is called a *dividing null hypothesis* (Cox, 1977; Bickel, 2011). The idea is that if the p value is low enough, then $\widehat{s} = \text{sign}\left(\widehat{\theta} - \theta_{H_0}\right)$ is a reasonable estimate of $s = \text{sign}(\theta - \theta_{H_0})$, where $\widehat{\theta}$ is an observed point estimate of θ and the function $\text{sign}(\bullet)$ has a value of 1 if its argument is positive, -1 if its argument is negative, and 0 otherwise. In that way, testing the null hypothesis that $\theta = \theta_{H_0}$ is used as an indirect method of deciding whether to claim that $s = \widehat{s}$.

A more direct way to make that decision would be to claim that $s = \widehat{s}$ only if it is sufficiently probable or, equivalently, if the *sign error* $s \neq \widehat{s}$ is sufficiently improbable. The sign error (Stephens, 2016) is also called a "Type III error" (Butler and Jones, 2018) and a "directional error" (Grandhi et al., 2019). The posterior probability of making a sign error given a two-sided p is

$$\Pr(s \neq \widehat{s} \mid P = p) = \begin{cases} \Pr(\theta > \theta_{H_0} \mid P = p) + \Pr(\theta = \theta_{H_0} \mid P = p) & \text{if } \widehat{\theta} < \theta_{H_0} \\ \Pr(\theta < \theta_{H_0} \mid P = p) + \Pr(\theta = \theta_{H_0} \mid P = p) & \text{if } \widehat{\theta} > \theta_{H_0} \end{cases}. \tag{4}$$

Under broadly applicable conditions, that is reasonably estimated by

$$\widehat{\Pr}(s \neq \widehat{s} \mid P = p) = (1 - v)\,\frac{p}{2} + v, \tag{5}$$

whenever v, the v value of equation (3), is a reasonable estimate of the $\Pr(\theta = \theta_{H_0} \mid P = p)$ in equation (2). The result is proved for all reasonable estimates of $\Pr(\theta = \theta_{H_0} \mid P = p)$ in Section 2.

The form of equation (5) represents a continuum between null hypothesis significance testing and conventional Bayesian testing. The frequentist practice of considering $\theta = \theta_{H_0}$ to be a dividing null hypothesis (Cox, 1977; Bickel, 2011) is recovered by setting $\Pr(\theta = \theta_{H_0}) = 0$, for in that case $v = 0$ and $\widehat{\Pr}(s \neq \widehat{s} \mid P = p) = p/2$, which is a one-sided p value. At the opposite extreme, the

traditional Bayesian practice of setting $\Pr(\theta = \theta_{H_0}) \geq 1/2$ often results in a v value that is much greater than the p value, in which case $\widehat{\Pr}(s \neq \widehat{s} | P = p) \approx v$. Choices of $\Pr(\theta = \theta_{H_0})$ between those frequentist and Bayesian extremes place $\widehat{\Pr}(s \neq \widehat{s} | P = p)$ within a continuum of values between $p/2$ and 1. For that reason, the easily interpreted estimate $\widehat{\Pr}(s \neq \widehat{s} | P = p)$ is a natural choice of a calibrated p value, as illustrated by example in Section 3. There, Figure 1 vividly portrays the Bayes-frequentist continuum.

The American Statistical Association's call to emphasize effect size estimation (Wasserstein and Lazar, 2016) does not necessarily warrant reporting conventional effect size estimates without modification (van den Bergh et al., 2019). In particular, a large effect size estimate can be misleading when a direction of the effect is too uncertain. To address that problem, Section 4 derives a simple calibration of the effect size estimate. The calibrated p value $\widehat{\Pr}(s \neq \widehat{s} | P = p)$ emerges as the degree of shrinkage.

Finally, implications for the debate and practice of testing null hypotheses are discussed in Section 5.

2 Estimating the local false sign rate of a single null hypothesis

For making connections to the literature and for succinctly deriving equation (5) regarding a test of the null hypothesis $\theta = \theta_{H_0}$, some terminology originally developed for testing multiple null hypotheses will prove useful. Since Efron et al. (2001) calls the $\Pr(\theta = \theta_{H_0} \mid P = p)$ of equation (2) the *local false discovery rate*, let $\text{LFDR} = \Pr(\theta = \theta_{H_0} \mid P = p)$; see Efron (2010) and Bickel (2019a) for expositions. Similarly, since Stephens (2016) calls the $\Pr(s \neq \widehat{s} | P = p)$ of equation (4) the *local false sign rate*, let $\text{LFSR} = \Pr(s \neq \widehat{s} | P = p)$.

As equation (4) suggests, to estimate LFSR of a single null hypothesis, we need not only $\widehat{\text{LFDR}}$, an estimate of LFDR, but also estimates of $\Pr(\theta \gtrless \theta_{H_0} \mid P = p)$. Seeing that

$$\begin{aligned}
\Pr(\theta \gtrless \theta_{H_0} \mid P = p) &= \Pr(\theta \gtrless \theta_{H_0}, \theta \neq \theta_{H_0} \mid P = p) \\
&= \Pr(\theta \neq \theta_{H_0} \mid P = p) \Pr(\theta \gtrless \theta_{H_0} \mid P = p, \theta \neq \theta_{H_0}) \\
&= (1 - \text{LFDR}) \Pr(\theta \gtrless \theta_{H_0} \mid P = p, \theta \neq \theta_{H_0}),
\end{aligned}$$

let $\widehat{\Pr}(\theta \gtrless \theta_{H_0} \mid P = p) = \left(1 - \widehat{\text{LFDR}}\right) p^{\lessgtr}$, where $p^{\lessgtr}$ is the estimate of $\Pr(\theta \gtrless \theta_{H_0} \mid P = p, \theta \neq \theta_{H_0})$

that is defined as a one-sided p value testing the null hypothesis that $\theta = \theta_{H_0}$ with $\theta \lesseqgtr \theta_{H_0}$ as the alternative hypothesis. From here on, the two-sided p value is $p = 2\min(p^<, p^>)$.

Estimating $\Pr(\theta \gtreqless \theta_{H_0} \mid P = p, \theta \neq \theta_{H_0})$ by $p^{\lesseqgtr}$ has both a Bayesian justification and a Fisherian justification. The Bayesian justification is that $p^{\lesseqgtr}$ is in many cases an approximation of a $\Pr(\theta \gtreqless \theta_{H_0} \mid P = p, \theta \neq \theta_{H_0})$ based on any member of a wide class of prior distributions that do not concentrate prior probability at θ_{H_0} or at any other point (Pratt, 1965; Casella and Berger, 1987). Setting $\Pr(\theta = \theta_{H_0}) > 0$ need not conflict with those priors since $\Pr(\theta \gtreqless \theta_{H_0} \mid P = p, \theta \neq \theta_{H_0})$, unlike $\Pr(\theta \gtreqless \theta_{H_0} \mid P = p)$, is conditional on $\theta \neq \theta_{H_0}$ (cf. Bickel, 2012b, 2018).

The Fisherian justification is that $p^{\lesseqgtr}$, as a fiducial probability or observed confidence level (Polansky, 2007) that $\theta \gtreqless \theta_{H_0}$ (Bickel, 2011), can serve as an *estimate* of a posterior probability that $\theta \gtreqless \theta_{H_0}$ even though, as many have noted (e.g., Grundy, 1956; Lindley, 1958; Evans, 2015, §3.6), it does not necessarily satisfy the properties of a Bayesian posterior probability. In the same way, many optimal point estimates can have values that are not possible for the parameters they estimate (Bickel, 2019b). That is why Wilkinson (1977, §6.2) considered fiducial probability as an *estimate* of a level of belief rather than as a level of belief. Similarly, confidence distributions, a modern development of fiducial distributions (Nadarajah et al., 2015), have been interpreted in terms of estimating θ (Singh et al., 2007; Xie and Singh, 2013) or an indicator of hypothesis truth (Bickel, 2012a).

Plugging the above estimates into equation (4) yields

$$\widehat{\text{LFSR}} = \begin{cases} \left(1 - \widehat{\text{LFDR}}\right) p^< + \widehat{\text{LFDR}} & \text{if } \widehat{\theta} < \theta_{H_0} \\ \left(1 - \widehat{\text{LFDR}}\right) p^> + \widehat{\text{LFDR}} & \text{if } \widehat{\theta} > \theta_{H_0} \end{cases}. \tag{6}$$

Theorem 1. *If* $\operatorname{sign}\left(\widehat{\theta} - \theta_{H_0}\right) = \operatorname{sign}(p^< - p^>)$*, then*

$$\widehat{\text{LFSR}} = \left(1 - \widehat{\text{LFDR}}\right) \frac{p}{2} + \widehat{\text{LFDR}}.$$

Proof. By equation (6), it is sufficient to prove that

$$p = \begin{cases} 2\,p^< & \text{if } \widehat{\theta} < \theta_{H_0} \\ 2\,p^> & \text{if } \widehat{\theta} > \theta_{H_0} \end{cases}.$$

Since the $\operatorname{sign}\left(\widehat{\theta} - \theta_{H_0}\right) = \operatorname{sign}(p^< - p^>)$ condition implies that $\widehat{\theta} < \theta_{H_0} \iff p^< < p^>$ and

$\widehat{\theta} > \theta_{H_0} \iff p^{>} < p^{<}$, it is enough to prove that

$$p = \begin{cases} 2\,p^{<} & \text{if } p^{<} < p^{>} \\ 2\,p^{>} & \text{if } p^{>} < p^{<} \end{cases},$$

which follows immediately from $p = 2\min\left(p^{<}, p^{>}\right)$. □

The sign $\left(\widehat{\theta} - \theta_{H_0}\right) = \text{sign}\left(p^{<} - p^{>}\right)$ condition for the theorem says the sign estimated by the parameter estimate agrees with the sign indicated by the one-sided p values. It holds in nearly all real situations.

3 Estimates of local false sign rates as calibrated p values

The estimate of the local false sign rate approaches a local false discovery rate or a one-sided p value, depending on the limiting conditions.

Corollary 1. *If* $\text{sign}\left(\widehat{\theta} - \theta_{H_0}\right) = \text{sign}\left(p^{<} - p^{>}\right)$*, then* $\lim_{p\to 0} \widehat{\text{LFSR}} = \widehat{\text{LFDR}}$ *and* $\lim_{\widehat{\Pr}(\theta=\theta_{H_0})\to 0} \widehat{\text{LFSR}} = p\,/2$*, where* $\widehat{\Pr}\left(\theta = \theta_{H_0}\right)$ *is the prior probability that yields* $\widehat{\text{LFDR}}$ *as the posterior probability.*

Proof. By Bayes's theorem, $\widehat{\text{LFDR}} \to 0$ as $\widehat{\Pr}\left(\theta = \theta_{H_0}\right) \to 0$. Both claims then follow from Theorem 1. □

Since $p\,/2 = \min\left(p^{<}, p^{>}\right)$, that result justifies calling $\widehat{\text{LFSR}}$ the $\widehat{\text{LFDR}}$-*calibrated* p *value* and accordingly denoting it by $p\left(\widehat{\text{LFDR}}\right)$ to stress its dependence on the choice of an estimate of LFDR.

Example 2. A simple option for $\widehat{\text{LFDR}}$ is v, the lower bound given in equation (3), with $\widehat{\Pr}\left(\theta = \theta_{H_0}\right)$ in place of $\Pr\left(\theta = \theta_{H_0}\right)$. Then we write the v-calibrated p value as $p\left(v\right)$.

The resulting Bayes-frequentist continuum is displayed as Figure 1, with traditional frequentism at the left end of each plot and traditional Bayesianism at the right. Figure 2 zooms in on three points in the continuum. ▲

Many other lower bounds on LFDR are available (e.g., Held and Ott, 2018, and references). But why estimate the LFDR with an estimate of a lower bound such as the v value (Example 2)? There are multiple reasons to accept the v value as an adequate estimate of the LFDR. First, as the Bayes factor can be lower than $\underline{B}$ (Held and Ott, 2018), which is the Bayes factor bound behind the v value, the v value is not necessarily a lower bound on LFDR. Second, $\underline{B}$ is close to estimated Bayes factors for many studies in epidemiology, genetics, and ecology (Bayarri et al., 2016, Fig. 3),

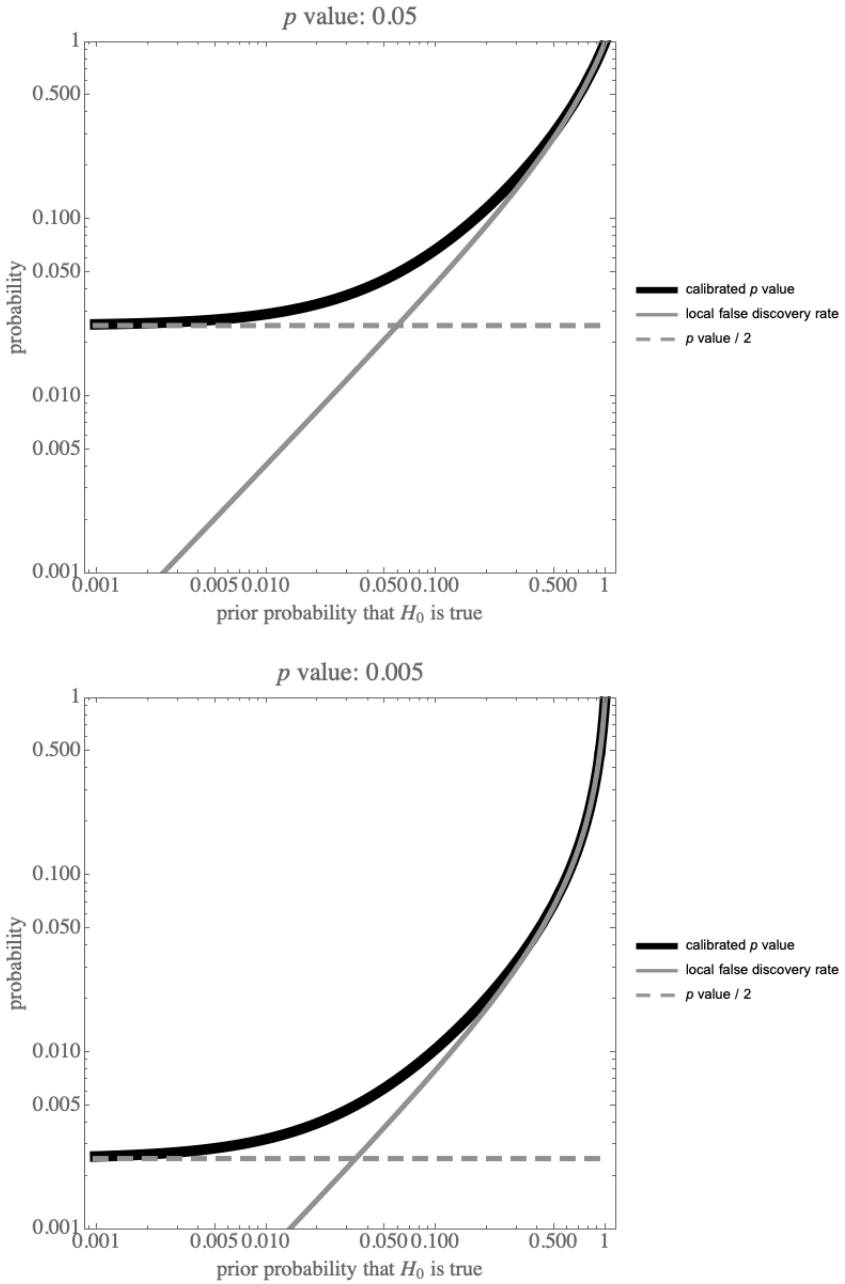

Figure 1: The three curves are $p\,(v)$, v, and $p\,/2$ as functions of $\Pr\,(\theta = \theta_{H_0})$. For both $p = 0.05$ and $p = 0.005$, the v-calibrated p value $p\,(v)$ approaches the one-sided p value $p\,/2$ as $\Pr\,(\theta = \theta_{H_0})$ decreases and approaches the estimated posterior probability v as $\Pr\,(\theta = \theta_{H_0})$ increases.

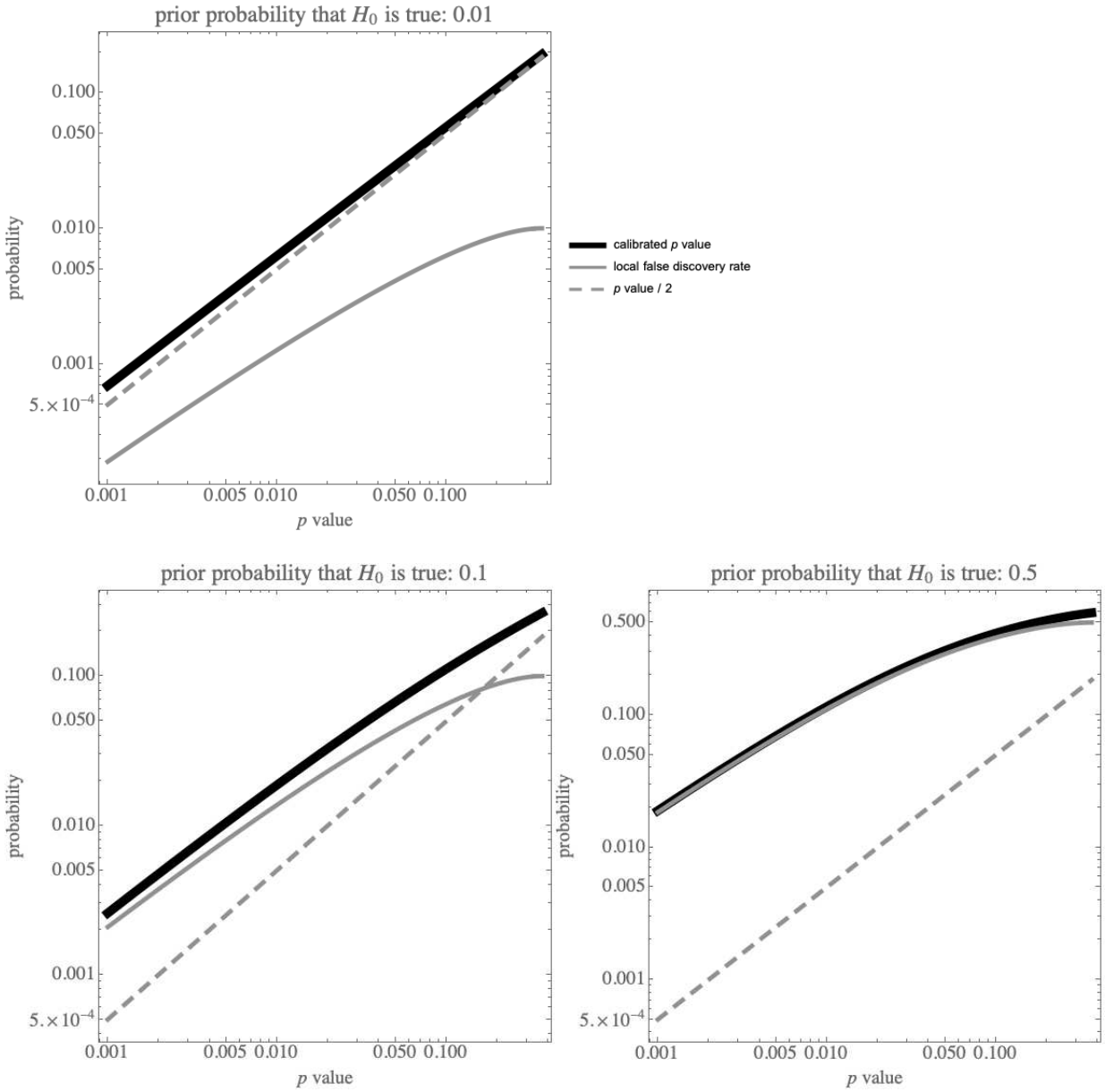

Figure 2: The three curves are $p\,(v)$, v, and $p\,/2$ as functions of p, the two-sided p value, for each of three prior probabilities: $\Pr\,(\theta = \theta_{H_0}) = 0.01, 0.1, 0.5$. In the plot corresponding most to traditional frequentism ($\Pr\,(\theta = \theta_{H_0}) = 0.01$), the v-calibrated p value $p\,(v)$ is close to $p\,/2$, a one-sided p value. In the plot corresponding most to traditional Bayesianism ($\Pr\,(\theta = \theta_{H_0}) = 0.5$), the v-calibrated p value $p\,(v)$ is close to v, the estimated posterior probability. The remaining plot ($\Pr\,(\theta = \theta_{H_0}) = 0.1$) shows a more interesting relationship between the v-calibrated p value, the estimated posterior probability, and the one-sided p value.

and the v value would be close in those cases to LFDR. Third, the v value is quantitatively similar to the following estimate of LFDR.

Example 3. Let z denote the probit transform of $p/2$; the probit function is implemented in R as `rnorm` and in Microsoft Excel as `norm.s.inv`. For $|z| \geq 1$, the *L value* is

$$L = \frac{1}{1 + 1/\widehat{B}},$$

where $\widehat{B} = 1.86\,|z|\,e^{-\frac{z^2}{2}}$ is the median-unbiased estimate of the Bayes factor assuming the probit transform of a one-sided p value is normal with mean 0 under $\theta \neq 0$ (Bickel, 2019a,d). (See Held and Ott (2016) for the maximum likelihood estimate under the same model and Pace and Salvan (1997) on the 0% confidence interval as a median-unbiased estimate.) Then $p(L)$ is the L-calibrated p value. It could be approximated by $p(v)$ since $p(L) \approx p(v)$, and the simplicity of $p(v)$ may make it more practical for general use (cf. Benjamin and Berger, 2019) than $p(L)$, which requires the probit transform. ▲

While the local false sign rate and local false discovery rate are posterior probabilities conditional on $P = p$, other posterior probabilities might serve as approximations.

Example 4. The *positive predictive value* $\Pr(\theta \neq \theta_{H_0} \mid P \leq \alpha)$ plays a key role in multiple papers related to the reproducibility crisis (e.g., Ioannidis, 2005; Button et al., 2013; Dreber et al., 2015; Wilson and Wixted, 2018). It is isomorphic to

$$\Pr(\theta = \theta_{H_0} \mid P \leq \alpha) = 1 - \Pr(\theta \neq \theta_{H_0} \mid P \leq \alpha),$$

which is known as the *false positive report probability* (Wacholder et al., 2004) and, in the multiple testing literature, as the *Bayesian false discovery rate* (Efron and Tibshirani, 2002) and the *nonlocal false discovery rate* (Bickel, 2013). An estimate of $\Pr(\theta = \theta_{H_0} \mid P \leq \alpha)$, such as the upper bound proposed by Bickel (2019c), is denoted by w and called a w *value* after Wacholder et al. (2004). Using it as an estimate of LFDR results in $p(w)$, the w-calibrated p value. However, w is highly biased as an estimate of LFDR when $\alpha = p$ (Colquhoun, 2017, 2019; Bickel and Rahal, 2019). ▲

4 Effect size estimation informed by local false sign rate estimation

If all relevant prior distributions were known, the Bayes-optimal estimate of the effect size θ under squared error loss would be its posterior mean,

$$\begin{aligned}
\mathrm{E}\left(\theta \middle| P = p\right) &= \Pr\left(s = \widehat{s} \middle| P = p\right) \mathrm{E}\left(\theta \middle| P = p, s = \widehat{s}\right) \\
&\quad + \Pr\left(s \neq \widehat{s}, \theta = \theta_{H_0} \mid P = p\right) \mathrm{E}\left(\theta \middle| P = p, s \neq \widehat{s}, \theta = \theta_{H_0}\right) \\
&\quad + \Pr\left(s \neq \widehat{s}, \theta \neq \theta_{H_0} \mid P = p\right) \mathrm{E}\left(\theta \middle| P = p, s \neq \widehat{s}, \theta \neq \theta_{H_0}\right) \\
&= \left(1 - \mathrm{LFSR}\right) \mathrm{E}\left(\theta \middle| P = p, s = \widehat{s}\right) + \left(\mathrm{LFDR}\right) \theta_{H_0} \\
&\quad + \left(\mathrm{LFSR} - \mathrm{LFDR}\right) \mathrm{E}\left(\theta \middle| P = p, s \neq \widehat{s}, \theta \neq \theta_{H_0}\right).
\end{aligned}$$

Without that knowledge, θ may instead be estimated by estimating $\mathrm{E}\left(\theta \middle| P = p\right)$.

In agreement with the $\widehat{\mathrm{LFSR}} = p\left(\widehat{\mathrm{LFDR}}\right)$ framework of Sections 2-3, $\mathrm{E}\left(\theta \middle| P = p\right)$ is estimated by the $\widehat{\mathrm{LFDR}}$-*calibrated effect size estimate*,

$$\begin{aligned}
\widehat{\theta}\left(\widehat{\mathrm{LFDR}}\right) &= \left(1 - p\left(\widehat{\mathrm{LFDR}}\right)\right) \widehat{\theta} + \left(\widehat{\mathrm{LFDR}}\right) \theta_{H_0} \\
&\quad + \left(p\left(\widehat{\mathrm{LFDR}}\right) - \widehat{\mathrm{LFDR}}\right) \theta_{H_0},
\end{aligned}$$

which uses $\widehat{\theta}$ to estimate $\mathrm{E}\left(\theta \middle| P = p, s = \widehat{s}\right)$ and θ_{H_0} to estimate $\mathrm{E}\left(\theta \middle| P = p, s \neq \widehat{s}, \theta \neq \theta_{H_0}\right)$. The latter estimate works best when θ would probably be close to θ_{H_0} conditional on a sign error. The calibrated effect size estimate simplifies to

$$\widehat{\theta}\left(\widehat{\mathrm{LFDR}}\right) = \left(1 - p\left(\widehat{\mathrm{LFDR}}\right)\right) \widehat{\theta} + p\left(\widehat{\mathrm{LFDR}}\right) \theta_{H_0}, \tag{7}$$

which reveals $p\left(\widehat{\mathrm{LFDR}}\right)$ as the degree to which $\widehat{\theta}$ is shrunk toward θ_{H_0}. The next result follows immediately from that and Corollary 1.

Corollary 2. *If* $\mathrm{sign}\left(\widehat{\theta} - \theta_{H_0}\right) = \mathrm{sign}\left(p^{<} - p^{>}\right)$, *then*

$$\lim_{p \to 0} \widehat{\theta}\left(\widehat{\mathrm{LFDR}}\right) = \left(1 - \widehat{\mathrm{LFDR}}\right) \widehat{\theta} + \widehat{\mathrm{LFDR}}\, \theta_{H_0}; \tag{8}$$

$$\lim_{\widehat{\Pr}\left(\theta = \theta_{H_0}\right) \to 0} \widehat{\theta}\left(\widehat{\mathrm{LFDR}}\right) = \left(1 - \frac{p}{2}\right) \widehat{\theta} + \frac{p}{2}\, \theta_{H_0}. \tag{9}$$

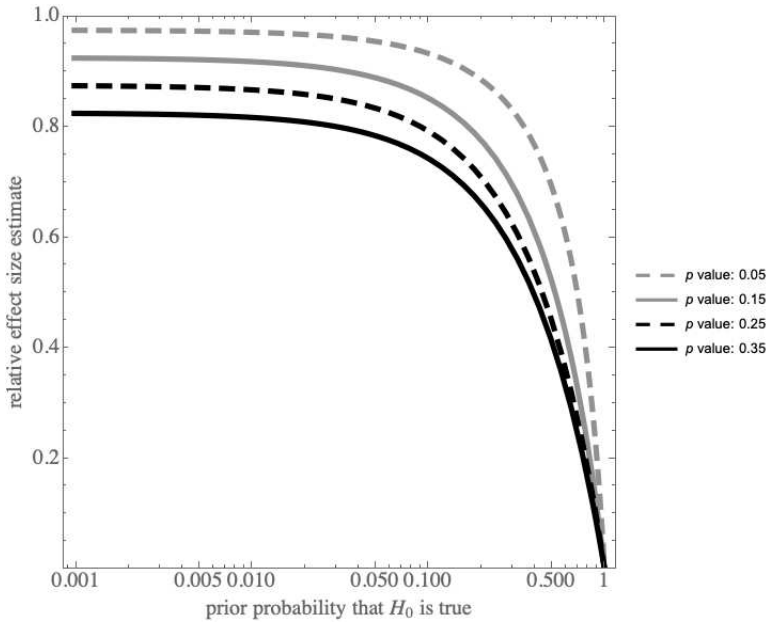

Figure 3: $\widehat{\theta}(v)/\widehat{\theta}$ as a function of $\widehat{\Pr}(\theta = 0)$ for $\theta_{H_0} = 0$ and $p = 0.05, 0.15, 0.25, 0.35$. The v-calibrated effect size estimate $\widehat{\theta}(v)$ is seen to shrink $\widehat{\theta}$ toward 0 as p or $\widehat{\Pr}(\theta = 0)$ increases.

The right-hand side of equation (8) has been used in multiple testing situations (e.g., Montazeri et al., 2010; Yanofsky and Bickel, 2010). Equation (9) records the effect of considering the local false sign rate even at the frequentist end of the Bayes-frequentist continuum.

An advantage of $\widehat{\theta}\left(\widehat{\text{LFDR}}\right)$ is that it shrinks $\widehat{\theta}$ toward θ_{H_0} more for higher p values without ever shrinking it all the way to θ_{H_0}, as seen in Figure 3. As a result, reporting calibrated effect size estimates could help prevent researchers from concluding that $\theta = \theta_{H_0}$ on the basic of a high p value.

5 Discussion

Imagine a world in which abstracts have v-calibrated effect size estimates and "p(v)=0.04," "p(v)=0.01," etc. in place of our world's uncalibrated estimates and "p<0.05." Adopting the local false sign rate estimate as a calibrated p value may focus current discussions about estimation and testing. The traditional Bayesian and frequentist positions would no longer be incommensurate paradigms or matters of upbringing and taste but rather opposite directions on the continuum determined by the prior probability of the null hypothesis (Figures 1-2). Going forward, debates would then concentrate on ways to estimate the prior probability for each field, data type, or other reference class (cf. Lakens et al., 2018; de Ruiter, 2019). Progress is already being made in measuring how the prior is

influenced by a field's risk tolerance (Wilson and Wixted, 2018), echoing the report that a demand for novelty leads to less reproducible results (Open Science Collaboration, 2015).

Even before a consensus is reached, statisticians can inform their collaborators of the impact of the prior probability on the local false sign rate estimate and help them determine adequate estimates of the prior for the data at hand. Estimates may be available in some cases from meta-analyses. For example, Benjamin et al. (2017) derived their infamous 0.005 significance threshold in part from meta-analyses suggesting $\widehat{\Pr}(\theta = \theta_{H_0}) = 10/11$ in psychology (Dreber et al., 2015; Johnson et al., 2017). The high value of that estimate reflects modeling assumptions that would in effect include values of θ that are close to θ_{H_0} with the null hypothesis rather than the alternative hypothesis. How close is close enough for inferential purposes may be a fruitful subject of future study and argument since it determines the calibrated p value through $\widehat{\Pr}(\theta = \theta_{H_0})$.

The difficulties involved in estimating prior probabilities may at times force us to retreat back to null hypothesis significance testing without any prior or to traditional Bayesian testing with the default 50% prior probability. The calibrated p value would then tell us what the estimated probability of making a sign error would be if the prior probability of the null hypothesis were actually 0% or 50%, respectively.

Acknowledgments

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

References

Bayarri, M., Benjamin, D. J., Berger, J. O., Sellke, T. M., 2016. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. Journal of Mathematical Psychology 72, 90–103.

Begley, C. G., Ioannidis, J. P., 2015. Reproducibility in science. Circulation Research 116 (1), 116–126.

Benjamin, D. J., Berger, J. O., 2019. Three recommendations for improving the use of p-values. The American Statistician 73 (sup1), 186–191.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R.,

Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., Johnson, V. E., 9 2017. Redefine statistical significance. Nature Human Behaviour, 1.

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

Bickel, D. R., 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., 2013. Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Statistical Applications in Genetics and Molecular Biology 12, 529–543.

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

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

Bickel, D. R., 2019b. 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., 2019c. Null hypothesis significance testing defended and calibrated by Bayesian model checking. The American Statistician, DOI: 10.1080/00031305.2019.1699443.
URL https://doi.org/10.1080/00031305.2019.1699443

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

Bickel, D. R., Rahal, A., 2019. Correcting false discovery rates for their bias toward false positives. Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1630432.

Butler, J. S., Jones, P., Apr 2018. Theoretical and empirical distributions of the p value. METRON 76 (1), 1–30.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., Munafò, M. R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14 (5), 365.

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.

Colquhoun, D., 2017. The reproducibility of research and the misinterpretation of p-values. Royal Society Open Science 4 (12), 171085.

Colquhoun, D., 2019. The false positive risk: A proposal concerning what to do about p-values. The American Statistician 73 (sup1), 192–201.

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

de Ruiter, J., Apr 2019. Redefine or justify? comments on the alpha debate. Psychonomic Bulletin & Review 26 (2), 430–433.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., 2015. Using prediction markets to estimate the reproducibility of scientific research. Proceedings of the National Academy of Sciences 112 (50), 15343–15347.

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

Efron, B., Tibshirani, R., 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, 70–86.

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.

Goodman, S. N., 06 1999. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. Annals of Internal Medicine 130 (12), 1005–1013.

Grandhi, A., Guo, W., Romano, J., 2019. Control of directional errors in fixed sequence multiple testing. Statistica Sinica 29 (2), 1047–1064.

Grundy, P. M., 1956. Fiducial distributions and prior distributions: An example in which the former cannot be associated with the latter. Journal of the Royal Statistical Society, Series B 18, 217–221.

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.

Hughes, B., 2018. Psychology in Crisis. Palgrave, London.

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.

Ioannidis, J. P., 2005. Why most published research findings are false. PLoS Medicine 2 (8), e124.

Johnson, V., Payne, R., Wang, T., Asher, A., Mandal, S., 2017. On the reproducibility of psychological science. Journal of the American Statistical Association 112 (517), 1–10.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., et al., 2018. Justify your alpha. Nature Human Behaviour 2 (3), 168.

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

Mayo, D. G., 2019. The ASA's p-value project: Why it's doing more harm than good (cont from 11/4/19). Web page, accessed 3 December 2019.
URL http://bit.ly/2LgXMKY

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

Montazeri, Z., Yanofsky, C. M., Bickel, D. R., 2010. Shrinkage estimation of effect sizes as an alternative to hypothesis testing followed by estimation in high-dimensional biology: Applications to differential gene expression. Statistical Applications in Genetics and Molecular Biology 9, 23.

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

Nieuwenhuis, S., Forstmann, B. U., Wagenmakers, E.-J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. Nature neuroscience 14 (9), 1105.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251).

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

Polansky, A. M., 2007. Observed Confidence Levels: Theory and Application. Chapman and Hall, New York.

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

Schachtman, N. A., 2019. Palavering about p-values. Web page, accessed 3 December 2019.
URL http://schachtmanlaw.com/palavering-about-p-values/

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

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.

Stephens, M., 10 2016. False discovery rates: a new deal. Biostatistics 18 (2), 275–294.

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., Wagenmakers, E.-J., Nov 2019. A cautionary note on estimating effect size. PsyArXiv, DOI: 10.31234/osf.io/h6pr8.
URL psyarxiv.com/h6pr8

Vovk, V. G., 1993. A logic of probability, with application to the foundations of statistics. Journal of the Royal Statistical Society: Series B (Methodological) 55 (2), 317–341.

Wacholder, S., Chanock, S., Garcia-Closas, M., Ghormli, L. E., Rothman, N., 2004. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. Journal of the National Cancer Institute 96, 434–442.

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

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.

Wilson, B. M., Wixted, J. T., 2018. The prior odds of testing a true effect in cognitive and social psychology. Advances in Methods and Practices in Psychological Science 1 (2), 186–197.

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

Yanofsky, C. M., Bickel, D. R., 2010. Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing. BMC Bioinformatics 11, art. 63.