Confidence distributions and empirical Bayes posterior distributions unified as distributions of evidential support

David R. Bickel

December 31, 2018

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

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

dbickel@uottawa.ca

Abstract

While empirical Bayes methods thrive in the presence of the hundreds or thousands of simultaneous hypothesis tests in genomics and other large-scale applications, significance tests and confidence intervals are considered more appropriate for small numbers of simultaneously tested hypotheses. Indeed, for fewer hypotheses or, more generally, fewer populations, there is more uncertainty in empirical Bayes methods of estimating the prior distribution. Confidence distributions have been used to propagate the uncertainty in the prior to empirical Bayes inference about a parameter, but only by combining a Bayesian posterior distribution with a confidence distribution, a probability distribution that encodes significance tests and confidence intervals. Combining distributions of both types has also been used to combine empirical Bayes methods and confidence intervals for estimating a parameter of interest. To clarify the foundational status of such combinations, the concept of an evidential model is proposed. In the framework of evidential models, both Bayesian posterior distributions and confidence distributions are degenerate special cases of evidential support distributions. Evidential support distributions, by quantifying the sufficiency of the data as evidence, leverage the strengths of Bayesian posterior distributions and confidence distributions for cases in which each type performs well and for cases benefiting from the combination of both. Evidential support distributions also address problems of bioequivalence, bounded parameters, and the lack of a unique confidence distribution.

Keywords: approximate confidence distribution; bioequivalence; bounded parameter; empirical Bayes methods; epistemic probability; evidential model; evidential support distribution; fiducial model averaging

1 Introduction

1.1 Empirical Bayes methods, confidence methods, and their evidential unification

Since the beginning of the century, the need to interpret genomics data has made unprecedented demands for innovations in multiple testing, leading to a resurgence of interest in empirical Bayes methods (e.g., Efron et al., 2001; Smyth, 2004; Qiu et al., 2005b,a; Scheid and Spang, 2005; Pan et al., 2008; Hong et al., 2009; Hwang et al., 2009; Ghosh, 2009; Muralidharan, 2010; Efron, 2015; Jiang and Yu, 2017; Karimnezhad and Bickel, 2018). On another front, to take back ground lost to fuller Bayesianism over the last few decades, a new frequentist offensive challenges its exclusive claim to posterior distributions. The long-discredited fiducial argument of Fisher has returned as various theories of confidence distributions (e.g., Schweder and Hjort, 2002; Singh et al., 2005; Polansky, 2007; Singh et al., 2007; Tian et al., 2011; Bityukov et al., 2011; Kim and Lindsay, 2011; Taraldsen and Lindqvist, 2018) and related priorless posterior distributions of the parameter of interest (e.g., Hannig et al., 2006; Hannig, 2009; Xiong and Mu, 2009; Gibson et al., 2011; Wang et al., 2012; Zhao et al., 2012; Balch, 2012; Martin and Liu, 2013; Bickel and Padilla, 2014; Bowater, 2017). Efron (2010), Nadarajah et al. (2015), and Schweder and Hjort (2016) provide informative expositions.

Unfortunately, neither comeback of frequentist ideas can subsume the other as a general approach to statistical inference. Without access to Bayes's theorem, pure confidence or fiducial theory falters in the presence of data relevant to so many hypotheses that a prior distribution can be reliably estimated (Robbins, 1985). On the other hand, traditional empirical Bayes methods only apply in the presence of such large-scale data sets.

Example 1. Rubin (1981) lists estimated test-score increase due to a training program for each of 8 educational sites involved in the study. The standard error of each estimate is also given, leading to a z score of $z_1 = 1.91$ for the first site. Thus, assuming z_1 was drawn from $\mathrm{N}(\mu_1, 1)$ with the test-score increase μ_1 unknown and testing the null hypothesis $H_0 : \mu_1 = 0$, the p value is greater than 0.05, and 0 would fall outside of the symmetric 95% confidence interval for μ_1. However, if it were known that $\mu_1 \in \{0, 2\}$, then

the Bayes factor would favor the alternative hypothesis $H_1 : \mu_1 = 2$ over the null hypothesis:

$$\frac{f_0\left(1.91\right)}{f_1\left(1.91\right)} = \frac{e^{-(1.91-0)^2/2}}{e^{-(1.91-2)^2/2}} = 0.16,$$

where f_0 and f_1 are the probability density functions under H_0 and H_1. In that case, the inference about the effect of the training program on the test score would depend on $\pi\left(0\right)$, the prior probability that $\mu_1 = 0$. The empirical Bayes method called Type II maximum likelihood (Good, 1966) calculates the posterior probability that $\mu_1 = 0$, which is

$$\pi\left(0|z_1\right) = \frac{\pi\left(0\right) f_0\left(z_1\right)}{\pi\left(0\right) f_0\left(z_1\right) + \left(1 - \pi\left(0\right)\right) f_1\left(z_1\right)} \tag{1}$$

according to Bayes's theorem, by replacing $\pi\left(0\right)$ with its maximum likelihood estimate based on the data from the $m = 8$ sites:

$$\widehat{\pi}_m\left(0\right) = \arg \sup_{\pi(0)\in[0,1]} \prod_{i=1}^{m} \pi\left(0\right) f_0\left(z_i\right) + \left(1 - \pi\left(0\right)\right) f_1\left(z_i\right).$$

If the study only had the first $m = 2$ sites, then the estimated posterior probability that $\mu_1 = 0$ would be $\widehat{\pi}\left(0|z_1\right) = 1.6 \times 10^{-8}$, representing a posterior odds of 7.8 orders of magnitude favoring H_1 over H_0. That is for all practical purposes certainty that $\mu_1 = 2$ rather than $\mu_1 = 0$ in spite of the unimpressive p value. That enormous overstatement of the evidence against the null hypothesis results from neglecting the variance in the estimated prior probability; the variance is considerable since data for only $m = 2$ sites were considered available. ▲

In the multiple testing literature, the posterior probability in equation (1) is known as the *local false discovery rate* (LFDR). Typical empirical Bayes LFDR estimates do not reflect the uncertainty in $\pi\left(0\right)$, in f_0, or in f_1 (Qiu et al., 2005b), motivating standard error estimates (Efron, 2007, §5) and confidence intervals (Scheid and Spang, 2005) for the LFDR.

Unfortunately, LFDR confidence intervals do not in themselves specify how to merge the uncertainty they convey about the LFDR with the uncertainty the LFDR conveys about H_0. To clearly assess how much support H_0 has from the evidence, the confidence intervals of the LFDR must somehow propagate uncertainty about the LFDR to uncertainty about whether the hypothesis is true. Confidence distributions can fill that gap in empirical Bayes theory.

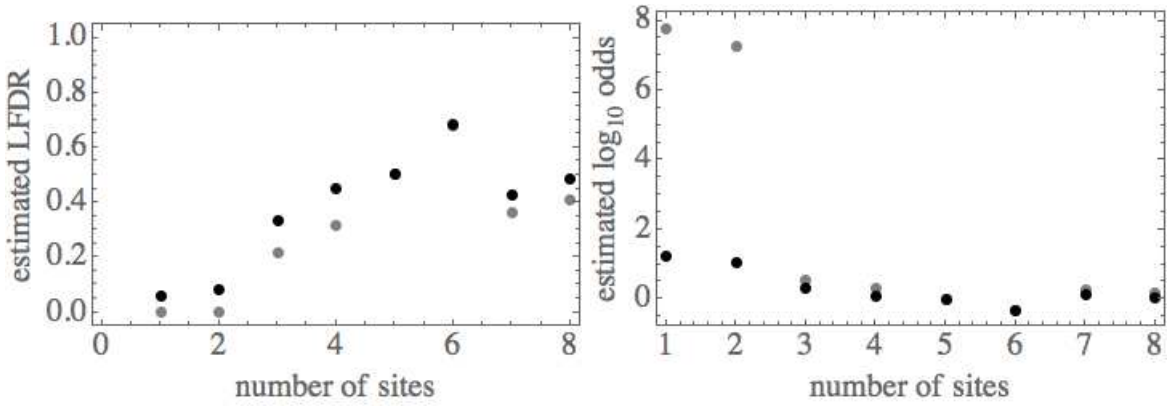

Figure 1: The estimated local false discovery rate at the first site (left plot) and the estimated posterior odds that there is an effect at the first site (right plot), with each estimate made on the basis of the first m sites, as functions of m, the number of sites. The gray dots in both plots are plug-in estimates, and the black dots are expected values of the LFDR (left plot) and the corresponding posterior odds (right plot). The estimation method is specified in Example 11.

Example 2. Returning to Example 1, the uncertainty in the prior probability can be accounted for by considering confidence intervals of that unknown probability at all levels of confidence. That induces a confidence distribution of prior probabilities and thus a distribution of posterior probabilities. Again assuming a study of 2 sites, the expectation value of the LFDR, a Bayesian posterior probability, is a non-Bayesian posterior probability that turns out to slightly favor the alternative hypothesis:

$$\text{Prob}\left(\mu_1 = 0; z_1\right) = E\left(\pi\left(0|z_1\right)\right) = 8.4\%. \tag{2}$$

That conclusion reflects all sources of uncertainty present according to the available evidence as encoded in the data, the model, and the confidence interval method. The use of confidence intervals alone or Type II maximum likelihood alone would have lead to very different and opposite conclusions (Example 1). As the number m of sites increases, the plug-in estimate tends to become closer to the expected LFDR (Figure 1), suggesting that the need to propagate the uncertainty in estimating the prior lessens as m increases. That trend mirrors the success of empirical Bayes methods for large-m data while indicating the need to supplement them with confidence distributions for small-m data. ▲

Example 2's expectation value is a non-Bayesian posterior probability equal to a Bayesian posterior probability integrated with respect to a confidence distribution. The point estimation or interval estimation

of a parameter of interest may instead or additionally call for the integration of a confidence distribution with respect to a Bayesian posterior distribution.

The standing taboo against both integrations stems largely from the fact that they lead to distributions on parameter space that in general are neither Bayesian posterior distributions nor confidence distributions. In fact, while confidence has a clear frequentist interpretation and Bayesian posterior probability has a clear interpretation as a limit of relative frequency when the prior distribution is essentially known from data, it is not clear how to interpret one type of probability integrated with respect to the other type. Subjective probability generalizes limits of relative frequencies but not confidence distributions, for they do not in general agree with Bayes's theorem unless the prior distribution is allowed to depend on the data.

The solution calls for a unified theory of marginalizing confidence distributions and Bayesian posterior distributions with respect to each other. Ideally, the theory would not only reduce to confidence theory and empirical Bayes estimation when each is appropriate but would also guide the analysis of data falling between those two extremes.

Toward that end, this paper develops the frequentist-Bayesian framework of Hill (1990) by defining the probability of a hypothesis about a parameter value as its evidential support, the extent to which a body of evidence supports the hypothesis. With the body of evidence including the model and any confidence region methods as well as the observed data, the distribution of evidential support generalizes confidence distributions and Bayesian posterior distributions. Leveraging the properties they hold as probability distributions leads to other distributions of evidential support. As a consequence, the class of evidential support distributions, unlike the class of confidence distributions, is closed to integration with respect to other distributions of the class.

1.2 Overview and applications

The modeling framework for evidential support distributions appears in Section 2, which defines the hierarchical evidential model as a generalization of the frequentist model and the Bayesian model. Section 3 presents settings that call for frequentist models in conjunction with Bayesian models.

Based on an evidential model, the evidential support distribution, defined in Section 4, both unifies confidence distributions and empirical Bayes distributions and provides an interpretation of the resulting probability distributions on parameter space. Examples in Section 5 are divided between those using condi-

tional confidence and those using hierarchical evidential models. Section 5.1 illustrates conditional confidence with applications to bioequivalence and parameter restrictions without priors. Section 5.2 highlights such applications of hierarchical evidential models as these:

- An explanation of how the evidential support for a null hypothesis is a generalization of the $E(\pi(0|z_1))$ mentioned in Example 2
- Estimates based on evidential support distributions that reduce to confidence intervals and other confidence regions when the null hypothesis has 0 prior probability
- The propagation of uncertainty involved in the lack of a unique confidence distribution given a parametric family of sampling distributions

The sense in which evidential support qualifies as a measure of evidence is discussed in Section 6. Specialized remarks on confidence and epistemic probability are postponed until Section 7.

2 Evidential models: frequentist, Bayesian, and hierarchical

Bayesian models (§2.1) and confidence models (§2.2) are the building blocks of hierarchical evidential models (§2.3).

2.1 Frequentist models, Bayesian models, and Bayesian evidential models

Given a *basic parameter* θ and a *nonbasic parameter* γ in sets Θ and Γ, let $f_{\theta,\gamma}$ denote a probability density function on a sample space $\mathcal{X}$. The terms "basic parameter" and "nonbasic parameter" replace the usual terms "parameter of interest" and "nuisance parameter" in order to define confidence intervals and other confidence regions for θ even when it is not the target of inference (Bickel and Padilla, 2014; Bickel, 2018a).

A *frequentist model* is a family $\{f_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}$, understood to be replaced by $\{f_\theta : \theta \in \Theta\}$ when there is no nonbasic parameter (Hill, 1990; Bickel, 2018a). In the context of a frequentist model, statements involving θ and γ hold for each $\theta \in \Theta$ and, if there is a nonbasic parameter, for each $\gamma \in \Gamma$.

With a prior probability density function π on $\Theta \times \Gamma$, a pair $(\pi, \{f_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\})$ is called a *Bayesian model* (Hill, 1990; Bickel, 2018a). The prior joint density of (θ, γ) is then $\pi(\theta, \gamma)$, and the prior marginal density of θ is $\pi(\theta) = E_\gamma(\pi(\theta, \gamma))$, which is $\int \pi(\theta, \gamma)\, d\gamma$ if π is relative to the Lebesgue measure. By

Bayes's theorem, the posterior probability densities for an observed sample x are $\pi(\theta, \gamma|x) \propto \pi(\theta, \gamma) f_{\theta,\gamma}(x)$ and $\pi(\theta|x) = E_\gamma(\pi(\theta, \gamma|x))$. The latter quantity defines the *Bayesian evidential model* as a function $(\theta, x) \mapsto \pi(\theta|x)$.

Hill (1990) defined a *statistical model* to be a set of Bayesian models defined with respect to the same parameter set, noting that it reduces to the case of a Bayesian model if the statistical model has a single member and to the case of a frequentist model if each member of the statistical model has the same frequentist model $\{f_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}$ but a different Dirac delta function $\delta_{\theta,\gamma}$, defined to have all of its probability mass at a parameter value $(\theta, \gamma) \in \Theta \times \Gamma$. In the latter case, the statistical model is

$$\{(\delta_{\theta',\gamma'}, \{f_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}) : \theta' \in \Theta, \gamma' \in \Gamma\}.$$

Hill (1990) argued that other cases call for empirical Bayes methods such as those that estimate a Bayesian model as a member of the statistical model by estimating the prior distribution.

2.2 Confidence curves, confidence densities, and confidence models

Under a frequentist model $\{f_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}$, a method of generating confidence intervals or other confidence sets is concisely represented as a function called a "confidence curve." Following Birnbaum (1961) and Blaker (2000), an *exact confidence curve* is a function $(\theta, x) \mapsto p(\theta; x)$ such that $p(\theta; X)$ is uniform on $[0, 1]$, where X is drawn from $f_{\theta,\gamma}$, that is,

$$\text{Prob}_{X \sim f_{\theta,\gamma}}(p(\theta; X) < \alpha) = \alpha \tag{3}$$

for every α between 0 and 1. The function $(\theta, x) \mapsto p(\theta; x)$ is called an *approximate confidence curve* if equation (3) holds up to some order of approximation (Bickel, 2018b; cf. Schweder and Hjort, 2016, p. 432). It follows that $p(\theta_0; x)$ is an observed p value testing $H_0 : \theta = \theta_0$ for any null hypothesis parameter value $\theta_0 \in \Theta$ and observed sample $x \in \mathcal{X}$. The curve gets its name from the fact that its inverse, $p^{-1}(\bullet; X)$, evaluated at $1 - \alpha$ for an α between 0 and 1, is an exact or approximate $100\%(1 - \alpha)$ confidence set, having

exactly or approximately $100\%\,(1-\alpha)$ frequentist coverage:

$$\text{Prob}_{X\sim f_{\theta,\gamma}}\left(p^{-1}\,(1-\alpha;X)\ni\theta\right)=1-\alpha.$$

In the special case that θ is a scalar, $p^{-1}\,(1-\alpha;X)$ is a $100\%\,(1-\alpha)$ confidence interval. Since every exact confidence curve is a degenerate case of an approximate confidence curve, the latter term will be used to encompass both concepts.

Let $(\theta;x)\mapsto c\,(\theta;x)$ denote a function such that $c\,(\bullet;x)$ is a probability density function on Θ for each $x\in\mathcal{X}$. An example in a Bayesian model is the posterior density function $\pi\,(\bullet|x)$ discussed in Section 2.1, but $c\,(\bullet;x)$ may be defined in a frequentist model without any prior distribution. If

$$\text{Prob}_{\vartheta\sim c(\bullet;x)}\left(\vartheta\in p^{-1}\,(1-\alpha;x)\right)=1-\alpha \tag{4}$$

for all α between 0 and 1 given an observed sample $x\in\mathcal{X}$, then $c\,(\bullet;x)$ is an *approximate confidence density function* corresponding to the approximate confidence curve $(\theta,x)\mapsto p\,(\theta;x)$ and the observation x (cf. Efron, 1993). According to equation (4), $c\,(\bullet;x)$ is the law of the random variable ϑ. An *approximate confidence distribution* corresponding to the approximate confidence curve $p\,(\bullet;\bullet)$ and the observation x is a probability measure that admits $c\,(\bullet;x)$ as an approximate confidence density function (Bickel, 2018b). Thus, every posterior distribution defined by applying Bayes's theorem to a probability matching prior is an example of an approximate confidence distribution (Bickel, 2012c, 2018b).

Since a frequentist model does not uniquely specify a procedure of generating confidence sets, a single frequentist model in general corresponds to multiple approximate confidence curves. (Pivotal models (Barnard, 1980, 1995, 1996) and structural models (Fraser, 1968, 1996) differ from frequentist models in that respect.) In addition, unless θ is a scalar (Bickel, 2012c), a single approximate confidence curve in general corresponds to multiple approximate confidence distributions. That is because equation (4) constrains $c\,(\bullet;x)$ without fully determining it in the case that θ is a vector (Bickel and Padilla, 2014).

Thus, in order to achieve the uniqueness of a Bayesian evidential model $\pi\,(\bullet|\bullet)$, a *frequentist evidential model* or *confidence model* corresponding to an approximate confidence curve $p\,(\bullet;\bullet)$ is defined as the function $(\theta;x)\mapsto c\,(\theta;x)$ such that, for every $x\in\mathcal{X}$, $c\,(\bullet;x)$ is an approximate confidence density function

corresponding to $p(\bullet;\bullet)$. For example, models that specify matching prior distributions (e.g., Helland, 2004, 2009) qualify as confidence models since they specify confidence methods in addition to frequentist models.

2.3 Evidential models and hierarchical evidential models

An *evidential model* on $\Theta \times \mathcal{X}$ is a function $(\theta, x) \mapsto \Pi(\theta; x)$ that is either a frequentist evidential model on $\Theta \times \mathcal{X}$ or a Bayesian evidential model on $\Theta \times \mathcal{X}$. Consider the case in which Θ is a set of other evidential models. Then any parameter value $\theta \in \Theta$ would be not only an index for $f_{\theta,\gamma}$ but also another evidential model. Such an evidential model is a *child* of $\Pi(\bullet;\bullet)$, which is the *parent* of the child and a *hierarchical evidential model*. The *descendants* of $\Pi(\bullet;\bullet)$ are its children, its children's children, etc.; the set of all descendants of $\Pi = \Pi(\bullet;\bullet)$ is denoted by $\mathcal{D}(\Pi)$.

3 Examples of hierarchical evidential models

These examples indicate several applications of hierarchical evidential models.

Example 3. In the simplest type of empirical Bayes estimation, an unknown prior distribution of some parameter θ_i, describing the ith of a finite number of populations, is estimated from a data set x consisting of multiple samples $x_1, x_2, \ldots$ Each x_i is considered as if drawn from a distribution f_{θ_i} representing the ith population, where θ_i is in turn drawn from the unknown prior, as in Example 1.

Since point estimates of the hyperparameter ϕ labeling the unknown prior π_ϕ fail to account for estimation uncertainty, Laird and Louis (1987) represented that uncertainty by a distribution of bootstrap estimates of ϕ derived from repeatedly resampling the data with replacement. That bootstrap distribution may be interpreted as approximating a confidence distribution for ϕ since bootstrap distributions are asymptotic confidence distributions (Singh et al., 2007).

More generally, for each ϕ in some hyperparameter space Φ, a Bayesian evidential model is $(\theta, x) \mapsto \pi_\phi(\theta|x)$, and the confidence model is $(\phi; x) \mapsto c(\phi; x)$ for some approximate confidence density function $c(\bullet; x)$, not necessarily approximated by bootstrapping. That confidence model is hierarchical since each ϕ is an index for a Bayesian evidential model. Example 2 is a special case. ▲

The multiple-population data structure of Example 3 and other empirical Bayes methods is not required

for a confidence model to be hierarchical.

Example 4. Consider a confidence model $(M, x) \mapsto c(M; x)$ on $\mathcal{M} \times \mathcal{X}$, where $\mathcal{M}$ is a set of real numbers that index evidential models on Θ. The evidential model $c(\bullet; \bullet)$ is hierarchical since every $M \in \mathcal{M}$ refers to another evidential model, either another confidence model or a Bayesian evidential model. ▲

While Examples 3-4 put a confidence model over other evidential models, Examples 5-6 instead put Bayesian evidential models over other evidential models.

Example 5. Suppose confidence intervals for a parameter θ_i like that of Example 3 would be appropriate if it were known that $\theta_i \neq 0$ but that there is an assumed or reliably estimated probability $\pi(0)$ that $\theta_i = 0$. That happens, for example, in genomics applications involving thousands of hypothesis tests, each corresponding to a sample from a different population (Bickel, 2012b). Bayes's theorem leads to the *local false discovery rate*, the posterior probability that $\theta = 0$:

$$\pi(0|x_i) = \frac{\pi(0) f_0(x_i)}{\pi(0) f_0(x_i) + \pi(1) f_1(x_i)}, \tag{5}$$

for x_i, the sample from the ith population, where $\pi(1) = 1 - \pi(0)$ and $f_0(x_i)$ and $f_1(x_i)$ are the probability densities of the observation x_i under $\theta = 0$ and $\theta \neq 0$, respectively. In that setting, data corresponding to the multiple hypothesis tests are used to estimate $\pi(0)$, f_1, and sometimes even f_0 (Efron, 2010). (The uncertainty involved in such estimation is ignored here for simplicity but may be represented by following Example 3; see Example 12.)

This example involves three evidential models. First, there is the Bayesian evidential model

$$(M_i, x_i) \mapsto \pi(M_i|x_i) = \begin{cases} \pi(0|x_i) & \text{if } M_i = 0 \\ 1 - \pi(0|x_i) & \text{if } M_i = 1 \end{cases}, \tag{6}$$

where the basic parameter space is $\mathcal{M} = \{0, 1\}$ and $M_i \in \mathcal{M}$. Second, $M_i = 1$ refers to the frequentist evidential model $(\theta_i; x_i) \mapsto c_1(\theta_i; x_i)$, where $c_1(\bullet; x_i)$ is the approximate confidence density function corresponding to the confidence intervals that would be appropriate if it were known that $\theta_i \neq 0$. Third, $M_i = 0$ points to another Bayesian evidential model, $(\theta_i, x_i) \mapsto \delta(\theta_i)$, where δ is the Dirac delta function, for under that model, the prior density function has all its mass at $\theta_i = 0$, and likewise for the posterior density

function defined by Bayes's theorem. Model (6) is hierarchical since each value of M_i is an index for another evidential model.

The hierarchical model generates point and interval estimates not only for data drawn from multiple populations, corresponding to multiple hypothesis tests, but also for data drawn from a single population, corresponding to a single hypothesis test (Bickel, 2012b). For an instance of the latter case, assume the Bayes factor $f_0(x_1)/f_1(x_1)$ is equal to a lower bound $B_0(x_1)$ that is a function of a p value according to one of the methods reviewed by (Held and Ott, 2018). Then the local false discovery rate is

$$\pi(0|x_1) = \frac{\pi(0) B_0(x_1)}{\pi(0) B_0(x_1) + \pi(1)} = \frac{1}{1 + \frac{\pi(1)}{\pi(0) B_0(x_1)}},$$

where $\pi(0)$ is assumed to be known, perhaps $\pi(0) = 1/2$ by symmetry or $\pi(0) = 10/11$, as suggested by meta-analyses (Benjamin et al., 2017). For point estimation of θ, its corresponding expected value according to the evidential support distribution is

$$E(\vartheta) = \text{Prob}(\vartheta = 0) E(\vartheta|\vartheta = 0) + \text{Prob}(\vartheta \neq 0) E(\vartheta|\vartheta \neq 0) = 0 + (1 - \pi(0|x_1)) \int \theta_1 c_1(\theta_1; x_i) d\theta_1.$$

▲

The next example has less of an empirical Bayes flavor.

Example 6. It often occurs in applications that more than one reasonable procedure for generating confidence sets corresponds to the same frequentist model. In such settings, the confidence model is not unique even though a single frequentist model is assumed. Let $\mathcal{M}$ denote a finite set of confidence models that correspond to that frequentist model. The prior probability mass function π on $\mathcal{M}$ reflects the weight given to each confidence method and thus to its confidence model. If there is no known reason to choose one reasonable confidence method over the other, then π assigns equal prior probability to each confidence model $M \in \mathcal{M}$. In general, because the likelihood function of M is constant on $\mathcal{M}$, the relevant Bayesian evidential model is $(M, x) \mapsto \pi(M|x) = \pi(M)$. That evidential model is hierarchical since every $M \in \mathcal{M}$ is another evidential model, in this case a frequentist evidential model. ▲

4 How much support a hypothesis has from the evidence

4.1 The evidence

A pair (x, Π) is a *body of evidence* or simply *the evidence* if $x \in \mathcal{X}$ is an observed sample and $\Pi = \Pi(\bullet; \bullet)$ is an evidential model on $\Theta \times \mathcal{X}$. It follows that a body of evidence has more information than what Birnbaum (1962) called "an instance of *statistical evidence*," a pair consisting of a frequentist model and an observed sample. Evans et al. (1986) point out that a model with more information than a frequentist model may falsify the premises from which Birnbaum (1962) derived the likelihood principle. Nevertheless, Bayesian models with priors not depending on their data distributions satisfy the likelihood principle, as do certain confidence models to the extent that their confidence sets and p values have little dependence on the choice of a confidence curve in the sense of Pierce and Peters (1994).

How much does the body of evidence support the inference that the data-generating value of θ is in some subset $\mathcal{H}$ of the parameter space? In other words, what is the degree of evidential support for the hypothesis that $\theta \in \mathcal{H}$? That is answered in the rest of this section by defining the concept of an evidential support distribution.

4.2 Conditional evidential support distributions given the evidence

To say a hypothesis "80% . . . supported by the data" is to leave 20% of the support for the statement that the hypothesis is false (Bickel, 2011). That way of speaking is captured by formalizing the evidential support for a hypothesis as a probability. More generally, the support from a body of evidence is probability mass with some distribution across the parameter set Θ. Each measurable subset of Θ then corresponds to a hypothesis of some amount of the support from the evidence.

A probability density function $\theta \mapsto s(\theta|x, \Pi)$ is a *conditional evidential support density function given a body of evidence* (x, Π) if $s(\bullet|x, \Pi) = \Pi(\bullet|x)$ in the case that $\Pi(\bullet; \bullet)$ is a Bayesian evidential model; in the case that $\Pi(\bullet; \bullet)$ is a frequentist evidential model,

$$\text{Prob}_{\vartheta \sim s(\bullet|x,\Pi)}(\vartheta \in \mathcal{H}) = \eta\left(\text{Prob}_{\vartheta \sim \Pi(\bullet;x)}(\vartheta \in \mathcal{H})\right) \tag{7}$$

for every measurable $\mathcal{H} \subset \Theta$, where η is a function on $[0, 1]$ with values in $[0, 1]$ that satisfies the *condition*

of universal calibration, that the same η applies for every frequentist evidential model.

The condition equating evidential support with posterior probability in the case of a Bayesian model formalizes the principle that a hypothesis is supported by the data and the model to the extent that they confer a high posterior probability to the hypothesis if the Bayesian model, including the prior, were known. It defines what it means for a hypothesis to have a certain amount of support from the body of evidence.

The conditions governing the case of the frequentist evidential model reflect the intuition that the amount of evidential support the hypothesis that θ lies in a $100\%\,(1-\alpha)$ confidence interval is a fixed function of the confidence level $100\%\,(1-\alpha)$. That intuition is reasonable when the confidence procedure is enough to assess the support from the evidence that includes x as the observation.

Theorem 1. *If a probability density function $\theta \mapsto s\,(\theta|x,\Pi)$ is a conditional evidential support density function given a body of evidence (x,Π), then $s\,(\bullet|x,\Pi) = \Pi\,(\bullet;x)$.*

Proof. Since the claim is given in the definition in the case that $\Pi\,(\bullet;\bullet)$ is a Bayesian evidential model, it suffices to prove it for the case that $\Pi\,(\bullet;\bullet)$ is a frequentist evidential model. Considering the Θ to be the real line, let $\mathfrak{P}$ denote a finite partition of Θ such that each member of the partition has equal probability according to $\Pi\,(\bullet;x)$. Then, since the total probability of $\Pi\,(\bullet;x)$ is 1,

$$\mathrm{Prob}_{\vartheta\sim\Pi(\bullet;x)}\,(\vartheta\in\mathcal{H}) = 1/\,|\mathfrak{P}|$$

for all $\mathcal{H}\in\mathfrak{P}$. Because the total probability of $s\,(\bullet|x,\Pi)$ must also be 1,

$$1/\,|\mathfrak{P}| = \mathrm{Prob}_{\vartheta\sim s(\bullet|x,\Pi)}\,(\vartheta\in\mathcal{H}) = \eta\left(\mathrm{Prob}_{\vartheta\sim\Pi(\bullet;x)}\,(\vartheta\in\mathcal{H})\right)$$

for all $\mathcal{H}\in\mathfrak{P}$, as per equation (7). Since that holds for finite partitions of arbitrarily small $1/\,|\mathfrak{P}|$, it follows that $\eta\,(1-\alpha) = 1-\alpha$ for all $\alpha\in[0,1]$. That can only be true if $s\,(\bullet|x,\Pi) = \Pi\,(\bullet;x)$. According to the condition of universal calibration, the same function η applies to all other frequentist evidential models, which implies that $s\,(\bullet|x,\Pi) = \Pi\,(\bullet;x)$ holds in general. □

While the result pertains to support from a body of evidence, it has consequences for epistemic probability (Remark 1). Section 6 discusses alternative ways to quantify the strength of evidence.

4.3 Evidential support distributions given the evidence

Let ψ denote a parameter in a set Ψ such that each basic parameter of an evidential model $\Pi(\bullet;\bullet)$ and the basic parameters of its descendants are subparameters of ψ. A probability density function $\psi \mapsto s(\psi|x)$ is a *basic evidential support density function* with respect to a body of evidence (x, Π) if it is the probability density function on Ψ of the *basic evidential support distribution*, the probability measure extended by $s(\bullet; x, \Pi)$ as a marginal probability density function and by each every conditional evidential support density function $s(\bullet|x, \Pi')$ for all $\Pi' \in \mathcal{D}(\Pi)$, where the extension is such that each $s(\bullet|x, \Pi')$ is a version of the conditional probability density function conditional on the event that the random variable of its parent's conditional evidential support density function is equal to Π'. Further, all probability measures derived from the basic evidential support distribution, including conditional probability distributions, laws of measurable functions, and the basic evidential support distribution, are *evidential support distributions.* Their probability density functions are called *evidential support density functions.*

5 Examples of evidential support distributions

These examples illustrate how to apply the definition of an evidential support distribution (§4.3) given an observation x.

5.1 Conditional confidence distribution given a subset of the parameter space

Before defining conditional confidence, it is motivated by example.

Example 7. Regulatory agencies often need to assess how much the evidence supports the hypothesis that a parameter value θ lies in $\theta' \pm \Delta$ for some $\theta' \in \mathbb{R}$ and $\Delta > 0$; a value common in bioequivalence studies is $\Delta = \ln(125\%)$ with $\exp(\theta')$ as the efficacy of a medical treatment. For the purpose of deciding whether to approve a new treatment or a genetically modified crop, estimates provided by companies with obvious conflicts of interest must be as objective as possible. The standard frequentist framework in effect enables conservative tests of the null hypotheses $H_0^{\text{equivalent}} : \theta \in [\theta' - \Delta, \theta' + \Delta]$, $H_0^{\text{lower}} : \theta < \theta' - \Delta$, and $H_0^{\text{higher}} : \theta > \theta' + \Delta$ (Wellek, 2003).

To use the approximate confidence curve corresponding to those p values to quantify the evidential

support for each of the three null hypotheses, consider the corresponding confidence density $c(\bullet;\bullet)$ as the frequentist evidential model and $(x, c(\bullet;\bullet))$ as the evidence. Thus, the evidential support density function is $s(\bullet|x, c(\bullet;\bullet)) = c(\bullet;x)$ according to Theorem 1. The integrals $\int_{-\infty}^{\theta'-\Delta} c(\theta;x)\,d\theta$, $\int_{\theta'-\Delta}^{\theta'+\Delta} c(\theta;x)\,d\theta$, and $\int_{\theta'+\Delta}^{\infty} c(\theta;x)\,d\theta$ are the evidential support probabilities of $H_0^{\text{equivalent}}$, H_0^{lower}, and H_0^{higher}; they are posterior probabilities that do not require any prior distribution. Since $c(\bullet;x)$ is a probability density function on the parameter space, regulators may also consider the evidential support for the hypothesis that the effect size is high given that it is not equivalent:

$$\text{Prob}_{\vartheta\sim c(\bullet;x)}(\vartheta > \theta' + \Delta | \vartheta \notin \theta' \pm \Delta) = \int_{\theta'+\Delta}^{\infty} c\left(\theta|\left[\theta'-\Delta, \theta'+\Delta\right]^c; x\right) d\theta$$

$$c\left(\theta|\left[\theta'-\Delta, \theta'+\Delta\right]^c; x\right) = \begin{cases} 0 & \text{if } \theta'-\Delta \le \theta \le \theta'+\Delta \\ \frac{c(\theta;x)}{1-\int_{\theta'-\Delta}^{\theta'+\Delta} c(t;x)dt} & \text{if } \theta \notin \theta' \pm \Delta \end{cases},$$

where $[\theta'-\Delta, \theta'+\Delta]^c$ is the complement of $\theta' \pm \Delta$. Singh et al. (2007) also compared the use of observed confidence levels to conventional methods of bioequivalence. ▲

More generally, the conditional probability density function

$$c(\theta|\mathcal{R};x) = \frac{c(\{\theta\}\cap\mathcal{R};x)}{\int_{\mathcal{R}} c(\theta;x)\,d\theta}$$

for a measurable *restriction set* $\mathcal{R} \in \Theta$ may be called an *approximate conditional confidence density function*. It qualifies as an evidential support density but not as an approximate confidence density. Conditional confidence distributions provide prior-free solutions to restricted parameter problems such as the following bounded parameter problem, in spite of Wilkinson (1977)'s dismissing such distributions as "Bayesian."

Example 8. Let $\mathcal{R}$ denote a bounded interval known to contain the value of θ, as is often appropriate in physics (Wang, 2007). The case of sampling a single observation from a normal distribution of unknown mean θ and known unit variance captures the essential features of the problem without obscuring them with a nuisance parameter and other complications (Fraser, 2011). In that case, if there were no parameter restriction, then $c(\bullet;x)$ would be an exact confidence density function, the probability density function of

$\mathrm{N}(x, 1)$, leading to

$$\mathrm{Prob}_{\vartheta \sim c(\bullet;x)}(\vartheta \leq \theta) = \int_{-\infty}^{\theta} c(t;x)\, dt = 1 - F(x-\theta),$$

for any real θ, where F is the standard normal CDF and x is the value of the observation. If a decay rate or neutrino mass is of interest, then $\mathcal{R} = [0, \infty[$ and

$$\begin{aligned}
\mathrm{Prob}_{\vartheta \sim c(\bullet;x)}(\vartheta \leq \theta | \vartheta \geq 0) &= \int_{-\infty}^{\theta} c(t|\,[0,\infty[\,;x)\, dt = \frac{\int_0^{\theta} c(t;x)\, dt}{\int_0^{\infty} c(t;x)\, dt} \\
&= \frac{\mathrm{Prob}_{\vartheta \sim c(\bullet;x)}(\vartheta < \theta) - \mathrm{Prob}_{\vartheta \sim c(\bullet;x)}(\vartheta < 0)}{1 - \mathrm{Prob}_{\vartheta \sim c(\bullet;x)}(\vartheta < 0)} \\
&= \frac{F(x) - F(x-\theta)}{F(x)} = 1 - \frac{F(x-\theta)}{F(x)}
\end{aligned}$$

for any $\theta \geq 0$.

Since $\mathrm{Prob}_{\vartheta \sim c(\bullet;x)}(\vartheta = 0 | \vartheta \geq 0) = \mathrm{Prob}_{\vartheta \sim c(\bullet;x)}(\vartheta \leq 0 | \vartheta \geq 0) = 1 - 1 = 0$, the frequentist evidential model ascribes zero restriction-conditional posterior probability to the boundary. In sharp contrast, the observed confidence at $\theta = 0$ is $1 - F(x)$ (Fraser, 2011), the single-sided p value with the null hypothesis at the boundary. In an investment application with θ_1 representing parameter bounded at 0, Schweder and Hjort (2016, §14.4) calculated the observed confidence at the boundary to be 90.03% and concluded, "We should be 90.03% confident that $\theta_1 = 0$." The conditional confidence approach avoids the problems of interpreting a p value as a posterior probability of a point null hypothesis, instead giving it zero non-Bayesian posterior probability.

However, the case of a non-zero posterior probability at a point in hypothesis space will arise as an evidential support probability under a Bayesian evidential model that assigns the point non-zero prior probability (Example 11 of Section 5.2). That would better represent the evidence when $\theta = 0$ is a viable possibility, as when θ is a radioactive decay rate, the mass of a neutrino (Mandelkern, 2002), an effect of extra-sensory perception (Bernardo, 2011), or the θ_1 in (Schweder and Hjort, 2016, §14.4). If, as is typical, the non-zero prior probability is unknown, then a frequentist evidential model can manage its uncertainty, as in Examples 11-12 of Section 5.2. ▲

5.2 Hypothesis support according to a hierarchical evidential model

The rest of the examples of evidential support distributions involve hierarchical evidential models from Section 3.

Example 9. Example 4 leads to the two-level hierarchy

$$M \sim c\left(\bullet; x\right)$$
$$\theta \sim \Pi_M\left(\bullet; x\right),$$

where M is the random index that points to the evidential model $\Pi_M\left(\bullet;\bullet\right)$. Thus, the evidential support density of each $\theta \in \Theta$ is

$$s\left(\theta; x\right) = \int_{M \in \mathcal{M}} c\left(M; x\right) \Pi_M\left(\theta; x\right) d\, M\,.$$

Since the expectation is with respect to the approximate confidence density $c\left(\bullet; x\right)$, the procedure is a form of fiducial model averaging (Bickel, 2015). It is called the *fiducial averaging of frequentist models* in the case that every $\Pi_M\left(\bullet;\bullet\right)$ is a frequentist evidential model and the *fiducial averaging of Bayesian models* in the case that every $\Pi_M\left(\bullet;\bullet\right)$ is a Bayesian evidential model (Bickel, 2018a).

In contrast, the next example explains a form of Bayesian model averaging of frequentist models or, more precisely, confidence models.

Example 10. Example 6 results in another two-level hierarchy,

$$M \sim \pi\left(\bullet|x\right) = \pi\left(\bullet\right)$$
$$\theta \sim c_M\left(\bullet; x\right),$$

where M is the random index referring to the confidence model $c_M\left(\bullet;\bullet\right)$. Thus, the evidential support density of each $\theta \in \Theta$ is

$$s\left(\theta; x\right) = \sum_{M \in \mathcal{M}} \pi\left(M\,|x\right) c_M\left(\theta; x\right) = \sum_{M \in \mathcal{M}} \pi\left(M\right) c_M\left(\theta; x\right).$$

In the special case of the uniform prior $\pi\left(\bullet\right) = 1/\left|\mathcal{M}\right|$, the evidential support distribution is the center of

mass (Paris, 1994) of the confidence distributions, a method Bickel (2012a,c) suggested for making inferences in the absence of a uniquely suitable confidence distribution. ▲

The following examples directly address the empirical Bayes concerns of Section 1.

Example 11. Example 3 generates an evidential support distribution that is a special case of Example 9's fiducial averaging of Bayesian models:

$$\begin{aligned}\phi &\sim c(\bullet; x)\\ M_i &\sim \pi_\phi(\bullet|x_i),\end{aligned}$$

with ϕ and M_i in place of the M and θ of Example 9. According to the evidential support distribution of (ϕ, M_i), the amount of evidential support for the ith null hypothesis is

$$\text{Prob}_{\phi\sim c(\bullet;x), M_i\sim\pi_\phi(\bullet|x_i)}(M_i = 0) = E_{\phi\sim c(\bullet;x)}\left(\text{Prob}_{M_i\sim\pi_\phi(\bullet|x_i)}(M_i = 0)\right) = E_{\phi\sim c(\bullet;x)}(\pi_\phi(0|x_i)),$$

which, as the confidence-averaged Bayesian posterior, is called the *fiducial Bayes probability* (Bickel, 2017).

The $E_{\phi\sim c(\bullet;x)}(\pi_\phi(0|x_1))$ denoted by $E(\pi(0|z_1))$ in Example 2 was calculated with $c(\bullet; x)$ as the confidence density function of the prior $\pi(0)$ that is derived from the first-order confidence curve of $\pi(0)$ based on the likelihood root statistic, following Bickel (2017, Example 3). $E_{\phi\sim c(\bullet;x)}(\pi_\phi(0|x_1))$ is also the expected LFDR with respect to an approximate confidence distribution of the LFDR (Remark 2). As suggested by Figure 1 of Example 2, when the estimation uncertainty incorporated into the expected LFDR is large, it can be much higher than the estimated LFDR. ▲

Example 12. Combining Example 3 with Example 5 yields the three-level hierarchy

$$\begin{aligned}\phi &\sim c(\bullet; x)\\ M_i &\sim \pi_\phi(\bullet|x_i)\\ \theta_i &\sim \chi(M_i = 0)\,\delta + \chi(M_i = 1)\,c_1(\bullet; x_i),\end{aligned}$$

where χ is the characteristic function and x consists of the samples $x_1, x_2, \ldots$. The resulting joint distribution of (ϕ, M_i, θ_i) is the evidential support distribution that generates set estimates of θ_i, including the

"propagated hierarchical set estimates" in Bickel (2017). By incorporating the uncertainty in ϕ, those set estimates differ for from the set estimates in Bickel (2012b), which are essentially based on Example 5's two-level model. ▲

6 Sufficiency of the evidence versus relevancy of the evidence

The two quantities most often considered as the strength of evidence in the statistics literature are the Bayes factor and the p value. They do not compete in the same league, for each measures a different sense of evidential strength.

The Bayes factor, as the ratio of the posterior odds to the prior odds, records the degree to which the data set, considered as the evidence, increases the support for one hypothesis over another (Lavine and Schervish, 1999). It measures the relevancy of the data set to the question of whether or not a hypothesis is true. In the special case that the hypotheses involved do not have nuisance parameters but correspond to distributions that may have generated the data, the Bayes factor is called the likelihood ratio. Other measures of the relevancy of the evidence include the relative belief ratio, the posterior probability divided by the prior probability (Evans, 2015), and the difference between posterior and prior probabilities (Kaye and Koehler, 2003).

The relevancy of the evidence to the truth of a hypothesis is not the same concept as the sufficiency of the evidence, which is the extent to which there is enough evidence to warrant a conclusion about the truth of the hypothesis (Kaye and Koehler, 2003). Thus, the posterior probability and posterior odds qualify as measures of the sufficiency of the evidence, provided that the prior probabilities are admitted as evidence (see Koehler, 2002). In addition, the p value is treated as the sufficiency of evidence whenever it is presented as if a low enough p value justifies the conclusion that the null hypothesis is false (e.g., Fraser et al., 2004), at least if there is sufficient power (Birnbaum, 1977). Similarly, Morgenthaler and Staudte (2012) argued for a function of a variance-stabilization p value with its standard error as a measure of evidence.

In the case of a scalar parameter of interest, a one-sided p value is equal to an observed confidence level, suggesting the confidence distribution's probability of a hypothesis as the sufficiency of the evidence supporting it. Whereas the probability of the hypothesis according to the confidence distribution is a measure of the sufficiency of the evidence in the absence of a prior (Bickel, 2011), the Bayesian posterior probability

is the sufficiency of the evidence in the presence of a prior. To handle other cases, both measures of the sufficiency of the evidence are generalized to evidential support in Section 4.

Example 13. In Examples 2 and 11, the sufficiency of the evidence for the hypothesis that the training program makes a difference at the first site is

$$\text{Prob}_{\phi\sim c(\bullet;x),M_1\sim\pi_\phi(\bullet|x_1)}\left(M_1=1\right)=E_{\phi\sim c(\bullet;x)}\left(\pi_\phi\left(1|x_1\right)\right)=1-E_{\phi\sim c(\bullet;x)}\left(\pi_\phi\left(0|x_1\right)\right),$$

the amount of evidential support for $M_1 = 1$, the first alternative hypothesis. The aptness of $E_{\phi\sim c(\bullet;x)}\left(\pi_\phi\left(1|x_1\right)\right)$ as the sufficiency of the evidence is clear given a loss function of the form

$$\ell_k\left(M_1,m_1\right)=\begin{cases}0 & \text{if } m_1=M_1\\ 1 & \text{if } m_1=0,M_1=1\\ k & \text{if } m_1=1,M_1=0,\end{cases} \tag{8}$$

where $m_1 = 0$ if it is concluded that $M_1 = 0$, $m_1 = 1$ if it is concluded that $M_1 = 1$, and $k > 0$. According to Remark 1, a decision maker whose evidence is confined to x and the hierarchical evidential model should decide on $\widehat{m}_1$, the value of m_1 that minimizes the expected loss

$$E_{\phi\sim c(\bullet;x),M_1\sim\pi_\phi(\bullet|x_1)}\left(\ell_k\left(M_1,m_1\right)\right)=\begin{cases}E_{\phi\sim c(\bullet;x)}\left(\pi_\phi\left(1|x_1\right)\right) & \text{if } m_1=0\\ kE_{\phi\sim c(\bullet;x)}\left(1-\pi_\phi\left(1|x_1\right)\right) & \text{if } m_1=1.\end{cases} \tag{9}$$

Thus, $\widehat{m}_1 = 0$ if $E_{\phi\sim c(\bullet;x)}\left(\pi_\phi\left(1|x_1\right)\right) < \left(1+k^{-1}\right)^{-1}$ but $\widehat{m}_1 = 1$ if $E_{\phi\sim c(\bullet;x)}\left(\pi_\phi\left(1|x_1\right)\right) > \left(1+k^{-1}\right)^{-1}$. In short, the evidence is sufficient to conclude that $M_1 = 1$ only if $E_{\phi\sim c(\bullet;x)}\left(\pi_\phi\left(1|x_1\right)\right)$ is high enough. In this example, the expected loss is the expectation value with respect to the derived confidence distribution of the loss (Remark 2). ▲

7 Remarks

Remark 1. The Bayesian-frequentist unification provided by Theorem 1 is an evidential analog of Wilkinson (1977, §3.1)'s epistemic concept of an "inferential probability" that is Bayesian posterior probability in the

presence of a prior but that is fiducial probability otherwise when confidence distributions are available (cf. Fisher, 1973; Zabell, 1992). Wilkinson (1977, §6.2) adds the qualification that whereas the former is a known degree of belief, the latter is an estimated degree of belief. Helland (2018) similarly proposes the concept of an epistemic process leading to confidence or Bayesian methods, depending on the availability of a prior. From outside of the statistics community, Franklin (2001) and Williamson (2013) instead consider confidence as a special case of logical probability and objective Bayesian probability, respectively. Without specifying how confidence applies in conjunction with Bayesian posterior probability, many others also interpret confidence not only as a limiting relative frequency but also as an epistemic probability (e.g., Hampel, 2006; Dempster, 2008; Bickel, 2012c; Schweder, 2018; Taraldsen and Lindqvist, 2018).

While stated in terms of impersonal support from evidence, Theorem 1 leads directly to an epistemic probability that encompasses confidence as well as Bayesian posterior probability. Evidential support for a hypothesis from a body of evidence is equal to the level of belief an agent should have for the hypothesis if the agent's body of knowledge is identical to the body of evidence. That epistemic probability is highly idealized, for a human agent's relevant body of knowledge is not limited to the body of evidence as defined in Section 4.1. For example, a statistician would be aware of limitations of the hierarchical evidential model that, with the sample x, constitutes the evidence.

This view of epistemic probability has a clear decision-theoretic interpretation: the ideal agent chooses the estimate or other action that minimizes expected loss with respect to the distribution of evidential support. Special cases include Example 13, the interpretation of confidence as a truth-value estimator (Bickel, 2012a), and empirical Bayes point estimates of the parameter of interest (Bickel, 2012b). The distribution of evidential support follows the evidential probability of Kyburg (1974, ch. 8; 1990, pp. 180, 231-234; 2003; 2006) in that it also prescribes decisions for an agent that has a specified body of knowledge.

Remark 2. Consider a measurable, strictly monotonic function of a basic parameter that has an approximate confidence distribution. Then the derived probability distribution of that function is itself an approximate confidence distribution.

For instance, the $\pi(0|x_i)$ of Examples 2 and 11 is strictly monotonic with $\pi(0)$ according to equation 5. It follows that the evidential support distribution of $\pi(0|x_i)$ induced by the approximate confidence distribution of ϕ, which is the random variable in place of $\pi(0)$, is an approximate confidence distribution that encodes confidence intervals with approximately correct frequentist coverage of $\pi(0|X_i)$.

Analogously for Example 13, the expected loss $E_{M_1 \sim \pi_\phi(\bullet|x_1)}\left(\ell_c\left(M_1, m_1\right)\right)$ is strictly monotonic with $\pi_\phi\left(1|x_1\right)$ according to equation (9). Therefore, the evidential support distribution of $E_{M_1 \sim \pi_\phi(\bullet|x_1)}\left(\ell_c\left(M_1, m_1\right)\right)$, also being derived from the confidence distribution of ϕ, is an approximate confidence distribution that encodes confidence intervals with approximately correct frequentist coverage of $E_{M_1 \sim \pi_\phi(\bullet|x_1)}\left(\ell_c\left(M_1, m_1\right)\right)$.

Were strict monotonicity not to hold, the evidential support distributions derived from the approximate confidence distributions would not in general be approximate confidence distributions.

Acknowledgments

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

References

Balch, M. S., 2012. Mathematical foundations for a theory of confidence structures. International Journal of Approximate Reasoning 53, 1003–1019.

Barnard, G., 1980. Pivotal inference and the Bayesian controversy. Trabajos de Estadistica Y de Investigacion Operativa 31, 295–318.

Barnard, G. A., 1995. Pivotal models and the fiducial argument. International Statistical Review / Revue Internationale de Statistique 63, 309–323.

Barnard, G. A., 1996. Corrigenda: Pivotal models and the fiducial argument. International Statistical Review / Revue Internationale de Statistique 64, 137–137.

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., 2012c. A frequentist framework of inductive reasoning. Sankhya A 74, 141–169.

Bickel, D. R., 2015. Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors. International Journal of Approximate Reasoning 66, 53–72.

Bickel, D. R., 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., 2018a. A note on fiducial model averaging as an alternative to checking Bayesian and frequentist models. Communications in Statistics - Theory and Methods 47, 3125–3137.

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

Bickel, D. R., Padilla, M., 2014. A prior-free framework of coherent inference and its derivation of simple shrinkage estimators. Journal of Statistical Planning and Inference 145, 204–221.

Birnbaum, A., 1961. Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. Journal of the American Statistical Association 56, pp. 246–249.

Birnbaum, A., 1962. On the foundations of statistical inference. Journal of the American Statistical Association 57, 269–305.

Birnbaum, A., 1977. The neyman-pearson theory as decision theory, and as inference theory; with a criticism of the lindley-savage argument for bayesian theory. Synthese 36 (1), 19–49.

Bityukov, S., Krasnikov, N., Nadarajah, S., Smirnova, V., 2011. Confidence distributions in statistical inference. AIP Conference Proceedings 1305, 346–353.

Blaker, H., 2000. Confidence curves and improved exact confidence intervals for discrete distributions. Canadian Journal of Statistics 28 (4), 783–798.

Bowater, R. J., 2017. A defence of subjective fiducial inference. AStA Advances in Statistical Analysis 101 (2), 177–197.

Dempster, A. P., 2008. The Dempster-Shafer calculus for statisticians. International Journal of Approximate Reasoning 48, 365–377.

Efron, B., 1993. Bayes and likelihood calculations from confidence intervals. Biometrika 80, 3–26.

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

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

Efron, B., 2015. Frequentist accuracy of Bayesian estimates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77 (3), 617–646.

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.

Evans, M. J., Fraser, D. A. S., Monette, G., 1986. On principles and arguments to likelihood. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 14, 181–194.

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

Franklin, J., 2001. Resurrecting logical probability. Erkenntnis 55, 277–305.

Fraser, D. A. S., 1968. The Structure of Inference. John Wiley, New York.

Fraser, D. A. S., 1996. Some remarks on pivotal models and the fiducial argument in relation to structural models. International Statistical Review / Revue Internationale de Statistique 64, 231–236.

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

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

Ghosh, D., 2009. Empirical Bayes methods for estimation and confidence intervals in high-dimensional problems. Statistica Sinica 19, 125–143.

Gibson, G. J., Streftaris, G., Zachary, S., 2011. Generalised data augmentation and posterior inferences. Journal of Statistical Planning and Inference 141, 156–171.

Good, I. J., 1966. How to Estimate Probabilities. IMA Journal of Applied Mathematics 2, 364–383.

Hampel, F., 2006. The proper fiducial argument. In: General Theory of Information Transfer and Combinatorics. Springer, pp. 512–526.

Hannig, J., 2009. On generalized fiducial inference. Statistica Sinica 19, 491–544.

Hannig, J., Iyer, H., Patterson, P., 2006. Fiducial generalized confidence intervals. Journal of the American Statistical Association 101, 254–269.

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

Helland, I. S., 2004. Statistical inference under symmetry. International Statistical Review 72, 409–422.

Helland, I. S., 2009. Steps Towards a Unified Basis for Scientific Models and Methods. World Scientific Publishing Company, Singapore.

Helland, I. S., 2018. Epistemic Processes. Springer, New York.

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

Hong, W.-J., Tibshirani, R., Chu, G., 2009. Local false discovery rate facilitates comparison of different microarray experiments. Nucleic Acids Research 37 (22), 7483–7497.

Hwang, J. T. G., Qiu, J., Zhao, Z., 2009. Empirical Bayes confidence intervals shrinking both means and variances. Journal of the Royal Statistical Society. Series B: Statistical Methodology 71, 265–285.

Jiang, W., Yu, W., 2017. Controlling the joint local false discovery rate is more powerful than meta-analysis methods in joint analysis of summary statistics from multiple genome-wide association studies. Bioinformatics 33 (4), 500–507.

Karimnezhad, A., Bickel, D. R., 2018. Incorporating prior knowledge about genetic variants into the analysis of genetic association data: An empirical Bayes approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics, DOI: 10.1109/TCBB.2018.2865420.
URL `https://ieeexplore.ieee.org/document/8436435/`

Kaye, D., Koehler, J., 2003. The misquantification of probative value. Law and Human Behavior 27 (6), 645–659.

Kim, D., Lindsay, B. G., 2011. Using confidence distribution sampling to visualize confidence sets. Statistica Sinica 21 (2), 923–948.

Koehler, J. J., 2002. When do courts think base rate statistics are relevant? Jurimetrics, 373–402.

Kyburg, H. E., 1974. The Logical Foundations of Statistical Inference. Reidel, Dordrecht ; Boston.

Kyburg, H. E., 2003. Are there degrees of belief? Journal of Applied Logic 1, 139–149.

Kyburg, H. E., 2006. Belief, evidence, and conditioning. Philosophy of Science 73, 42–65.

Kyburg, H. E. J., 1990. Science and Reason. Oxford University Press, New York.

Laird, N. M., Louis, T. A., 1987. Empirical Bayes confidence intervals based on bootstrap samples. Journal of the American Statistical Association 82, 739–750.

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

Mandelkern, M., 2002. Setting confidence intervals for bounded parameters. Statistical Science 17, 149–159.

Martin, R., Liu, C., 2013. Inferential Models: A Framework for Prior-Free Posterior Probabilistic Inference. Journal of the American Statistical Association 108 (501), 301–313.

Morgenthaler, S., Staudte, R. G., 2012. Advantages of Variance Stabilization. Scandinavian Journal of Statistics 39 (4), 714–728.

Muralidharan, O., 2010. An empirical Bayes mixture method for effect size and false discovery rate estimation. Annals of Applied Statistics 4, 422–438.

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

Pan, W., Jeong, K., Xie, Y., Khodursky, A., 2008. A nonparametric empirical Bayes approach to joint modeling of multiple sources of genomic data. Statistica Sinica 18, 709–729.

Paris, J. B., 1994. The Uncertain Reasoner's Companion: A Mathematical Perspective. Cambridge University Press, New York.

Pierce, D. A., Peters, D., 1994. Higher-order asymptotics and the likelihood principle: One-parameter models. Biometrika 81 (1), 1–10.

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

Qiu, X., Brooks, A., Klebanov, L., Yakovlev, A., 2005a. The effects of normalization on the correlation structure of microarray data. BMC Bioinformatics 6 (1), 120.

Qiu, X., Klebanov, L., Yakovlev, A., 2005b. Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. Statistical Applications in Genetics and Molecular Biology 4, art. 34.

Robbins, H., 1985. The empirical Bayes approach to statistical decision problems. In: Herbert Robbins Selected Papers. Springer, New York, pp. 49–68.

Rubin, D. B., 1981. Estimation in parallel randomized experiments. Journal of Educational Statistics 6, pp. 377–401.

Scheid, S., Spang, R., 2005. Twilight; a bioconductor package for estimating the local false discovery rate. Bioinformatics 21, 2921–2922.

Schweder, T., 2018. Confidence is epistemic probability for empirical science. Journal of Statistical Planning and Inference 195, 116–125.

Schweder, T., Hjort, N., 2016. Confidence, Likelihood, Probability. 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.

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.

Smyth, G. K., 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3.

Taraldsen, G., Lindqvist, B. H., 2018. Conditional fiducial models. Journal of Statistical Planning and Inference 195, 141–152.

Tian, L., Wang, R., Cai, T., Wei, L.-J., 2011. The highest confidence density region and its usage for joint inferences about constrained parameters. Biometrics 67 (2), 604–10.

Wang, C., Hannig, J., Iyer, H. K., 2012. Fiducial prediction intervals. Journal of Statistical Planning and Inference 142 (7), 1980–1990.

Wang, H., 2007. Modified p-values for one-sided testing in restricted parameter spaces. Statistics and Probability Letters 77, 625–631.

Wellek, S., 2003. Testing Statistical Hypotheses of Equivalence. Chapman and Hall, London.

Wilkinson, G. N., 1977. On resolving the controversy in statistical inference. JRSS B 39, 119–144.

Williamson, J., 2013. Why frequentists and Bayesians need each other. Erkenntnis 78 (2), 293–318.

Xiong, S., Mu, W., 2009. On construction of asymptotically correct confidence intervals. Journal of Statistical Planning and Inference 139 (4), 1394–1404.

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

Zhao, S., Xu, X., Ding, X., 2012. Fiducial inference under nonparametric situations. Journal of Statistical Planning and Inference 142, 2779 – 2798.