A Bayes Factor Framework for Unified Parameter Estimation and Hypothesis Testing

Samuel Pawel
Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science (CRS)
University of Zurich
E-mail: samuel.pawel@uzh.ch

Working paper version February 21, 2024

Abstract

The Bayes factor – the relative likelihood of the data under two competing hypotheses – is a natural measure of statistical evidence or support for one hypothesis over the other. Here we show how Bayes factors can also be used for parameter estimation. The key idea is to consider the Bayes factor as a function of the parameter value under the null hypothesis. This 'Bayes factor function' is then inverted to obtain point estimates ('maximum evidence estimates') and interval estimates ('support intervals'), similar to how *P*-value functions are inverted to obtain point estimates and confidence intervals. This provides data analysts with a unified inference framework for hypothesis testing and parameter estimation, as Bayes factors (for any tested parameter), support intervals (at any level of interest), and point estimates can be easily read off from a plot of the Bayes factor function. This approach shares similarities but is also distinct from conventional Bayesian and frequentist inference approaches. For instance, it uses the Bayesian evidence calculus, but without synthesizing data and prior, or it defines statistical evidence in terms of relative likelihoods, but also includes a natural way to deal with nuisance parameters. Applications to several real-world examples illustrate how our framework is of practical value to data analysts who aim to make quantitative inferences.

Keywords: Bayesian inference, integrated likelihood, meta-analysis, nuisance parameters, replication studies, support interval

1 Introduction

A universal problem in data analysis is making inferences about unknown parameters of a statistical model based on observed data. In practice, data analysts are often interested in two tasks: (i) estimating the parameters (i.e., finding the most plausible value or a range of plausible values based on the observed data) and (ii) testing hypotheses related to them (i.e., using the observed data to quantify the evidence that the parameter takes a certain value). While these tasks may seem different at first, there are several statistical concepts that provide a link between the two.

In frequentist statistics, there is a duality between parameter estimation and hypothesis testing as *P*-values, confidence intervals, and point estimates correspond in the sense that the *P*-value for

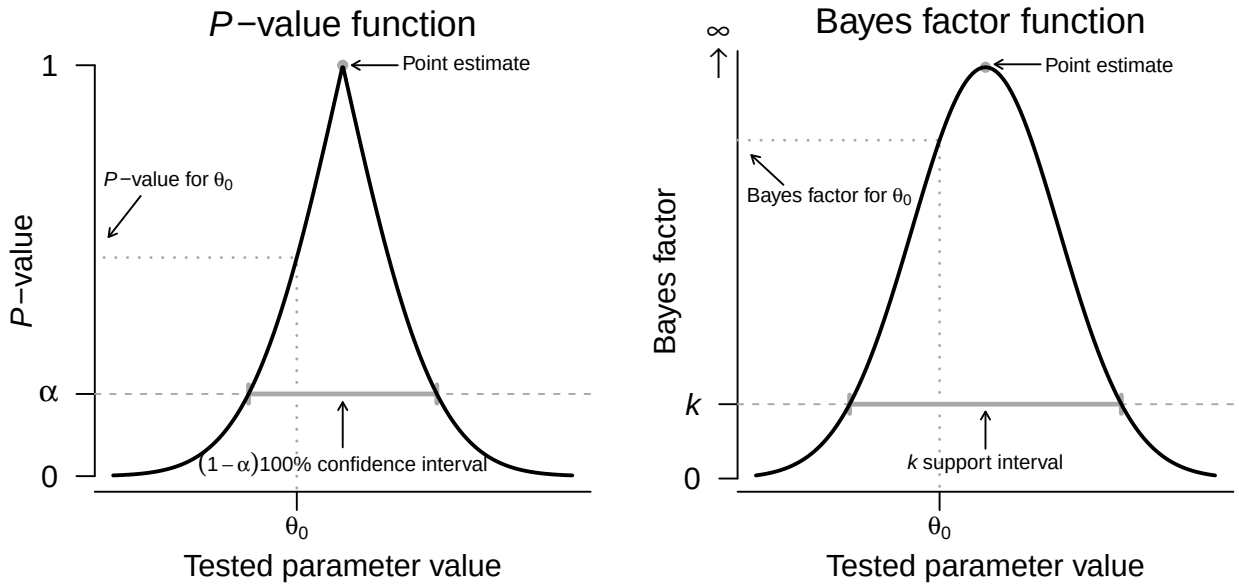

Figure 1: Examples of *P*-value functions and Bayes factor functions. *P*-value are two-sided. Bayes factors are oriented in favor of the tested parameter value over a specified alternative hypothesis (i.e., a higher Bayes factor indicates higher support for the parameter value over the alternative).

a tested parameter value is less than α if the $(1-\alpha)100\%$ confidence interval excludes that parameter value, and that the (two-sided) *P*-value is largest when the tested parameter value is the point estimate. The *P-value function* – the *P*-value viewed as a function of the tested parameter (for an overview see e.g., Bender et al., 2005; Fraser, 2019) – provides a link between these concepts. One may alternatively look at closely related quantities: One minus the two-sided *P*-value function known as *confidence curve* (Cox, 1958; Birnbaum, 1961), one minus the one-sided *P*-value function known as *confidence distribution*, or its derivative known as *confidence density* (Xie and Singh, 2013; Schweder and Hjort, 2016). A visualization of the *P*-value function, such as shown in the left plot in Figure 1, provides the observer with a wealth of information, as *P*-values (for any tested parameter), confidence intervals (at any level of interest), and point estimates can be easily read off. As such, *P*-value functions and their relatives have been deemed important measures to address common misinterpretations and misuses of *P*-values and confidence intervals (see e.g., Greenland et al., 2016; Infanger and Schmidt-Trucksäss, 2019; Rafi and Greenland, 2020; Marschner, 2024, among others).

In Bayesian statistics, the posterior distribution of the unknown parameter plays a similar role to the *P*-value function, since point estimates (e.g., posterior modes, medians, or means), credible intervals, and posterior probabilities of hypotheses can all be derived from it. The posterior provides a synthesis of the data and the prior distribution, which can be seen as an advantage but also as a challenge in the absence of prior knowledge. In particular, for testing of hypotheses, it can be difficult to specify prior probabilities such as '$\Pr(\text{the treatment effect is absent})$' and '$\Pr(\text{the treatment effect is present})$'. One approach to address this, is to report the *Bayes factor* (Jeffreys, 1939; Good, 1958; Kass and Raftery, 1995), i.e., the updating factor of the prior to posterior odds of two hypotheses. As such, Bayes factors allow data analysts to evaluate the relative evidence for two hypotheses without depending on the prior probabilities of the hypotheses themselves; for example, a Bayes factor can quantify the evidence for the presence or absence of a treatment effect

without having to assign prior probabilities to these hypotheses (although one still has to specify a prior for the parameter under the alternative, which is challenging in itself). However, the use of Bayes factors comes at the cost of lacking an overarching concept, such as a P-value function or posterior distribution, that can provide data analyst with a coherent set of point and interval estimates. In practice, data analysts who wish to perform hypothesis testing with Bayes factors but also parameter estimation are therefore faced with a dilemma; they can either supply their Bayes factors with a posterior distribution conditional on one hypothesis being true (e.g., the posterior of a treatment effect, assuming the effect is indeed present), which can lead to contradictory conclusions with the Bayes factor (for examples, see Stone, 1997; Wagenmakers et al., 2020), or they can assign prior probabilities to the tested hypotheses and report a posterior averaged over both hypotheses (Campbell and Gustafson, 2022), but this requires specification of prior probabilities which is highly controversial and the reason why the Bayes factor was reported in the first place rather than the posterior probabilities of the hypotheses.

Our goal is therefore to propose a unifying framework for estimation and hypothesis testing based on Bayes factors, building on ideas already hinted at in earlier work (Pawel et al., 2023b) and also closely related to the 'Bayes factor surface' (Fowlie, 2024) and 'K ratio' (Afzal et al., 2023) approaches recently proposed in the physics community. The idea is the same as for the P-value function; we consider the Bayes factor as a function of the tested parameter. We then use this *Bayes factor function*[1] to derive point estimates, interval estimates, and Bayes factors (as shown in the right plot in Figure 1). Our framework builds on the recently proposed Bayesian support interval (Wagenmakers et al., 2020; Pawel et al., 2023b) and extends it with the novel concept of point estimation based on Bayes factors. We call the resulting estimate the *maximum evidence estimate* (MEE) – the parameter value that receives the most evidential support from the data over a specified alternative hypothesis. This provides data analysts with a unified framework for statistical inference centred around the Bayes factor.

This paper is structured as follows. In the following (Section 2), we introduce the general theory of Bayes factors, support sets, and maximum evidence estimates. We then discuss their connection to other approaches to statistical inference (Section 3). Various real data examples in Section 4 illustrate properties of the Bayes factor framework. We conclude with a discussion of the advantages, limitations, and opportunities for future research (Section 5).

2 Bayes factor function inference

Suppose we observe data y with an assumed distribution with density/probability mass function $p(y \mid \theta, \psi)$ that depends on parameters $\theta \in \Theta$ and $\psi \in \Psi$, with θ being the focus parameters and ψ being possible nuisance parameters. Consider two hypotheses, the null hypothesis $H_0 \colon \theta = \theta_0$ postulating that θ takes a certain value θ_0 and the alternative hypothesis $H_1 \colon \theta \neq \theta_0$ postulating that θ takes a different value. A natural measure of relative evidence for the two hypotheses is the Bayes

[1] Our conception of Bayes factor functions differs from the Bayes factor functions introduced by Johnson et al. (2023). The latter are Bayes factors viewed as a function of a hyperparameter of the prior *under the alternative hypothesis* but for a fixed null hypothesis. We acknowledge that introducing another concept with the same name may be confusing, but we think that Bayes factor function is the most appropriate name, in analogy to P-value function. An alternative might be to call our concept 'support curve' in analogy to the confidence curve, but we think Bayes factor function is more appropriate.

factor (Jeffreys, 1939; Good, 1958; Kass and Raftery, 1995), the data-based updating factor of the prior odds of the hypotheses to their posterior odds

$$\text{BF}_{01}(y; \theta_0) = \frac{p(H_0 \mid y)}{p(H_1 \mid y)} \Big/ \frac{p(H_0)}{p(H_1)} \tag{1a}$$

$$= \frac{p(y \mid H_0)}{p(y \mid H_1)} \tag{1b}$$

$$= \frac{\int_\Psi p(y \mid \theta_0, \psi) p(\psi \mid H_0) \, \mathrm{d}\psi}{\int_\Theta \int_\Psi p(y \mid \theta, \psi) \, p(\theta, \psi \mid H_1) \, \mathrm{d}\psi \, \mathrm{d}\theta} \tag{1c}$$

with $p(\theta, \psi \mid H_1)$ denoting the prior assigned to the parameters under H_1 and $p(\psi \mid H_0)$ the prior assigned to the nuisance parameters under H_0.

As (1a) shows, the Bayes factor represents the data-based core of the Bayesian belief calculus. It remains useful even if one rejects the idea of assigning probabilities to H_0 and H_1, since this is not necessary (Goodman, 1999). The alternative expression of the Bayes factor in equation (1b) shows that this update is dictated by the relative predictive accuracy of the two hypotheses. That is, the posterior odds of the null hypothesis H_0 increase if it outperforms the competing alternative hypothesis H_1 in predicting the data y, and vice versa (Good, 1952; Gneiting and Raftery, 2007). Finally, the last equation (1c) shows how the Bayes factor can be calculated, i.e., by dividing the likelihood of y under the null value θ_0 (possibly marginalized over the prior of ψ under H_0) by the likelihood of y marginalized over the prior of θ (and possibly ψ) under H_1. The priors for θ and ψ may also be point priors, in which case the Bayes factor reduces to a likelihood ratio.

The idea now is to consider the Bayes factor (1) as a function of θ_0, that is, to vary the tested parameter value (the point null hypothesis $H_0 \colon \theta = \theta_0$) in order to assess the support for different parameter values over the alternative H_1, see the right plot in Figure 1 for an example. Like the P-value function, this *Bayes factor function* (BFF) helps to address cognitive challenges with inferential statistics (Greenland, 2017). For example, it shifts the focus of inference from testing a single privileged null hypothesis (e.g., the hypothesis that there is no treatment effect) to an entire continuum of hypotheses. By looking at the BFF, data analysts can then identify hypotheses that receive equal or even less support from the data than the privileged one; for example, a parameter value indicating a very large treatment effect may receive equal support as the value of no treatment effect (sometimes called 'counternull', see Rosenthal and Rubin, 1994).

For one- or two-dimensional focus parameters θ, the BFF can be plotted as a curve or surface, respectively, so that the relative support for parameter values can be visually assessed. For higher dimensional focus parameters, this becomes more difficult and the BFF may need to be summarized in some way, which we discuss in the following.

2.1 Support sets

The BFF can be used to obtain *support sets* (Wagenmakers et al., 2020; Pawel et al., 2023b) which are set-valued estimates for θ based on inverting the Bayes factor (1) similar to how P-value functions

are inverted to obtain confidence sets. Specifically, a support set at support level $k > 0$ is defined by

$$S_k = \{\theta_0 : \text{BF}_{01}(y; \theta_0) \geq k\}$$

that is, the parameter values for which the Bayes factor indicates at least evidence of level k over the specified alternative. In practice, a k support set (typically an interval) is obtained from 'cutting' the BFF at k and taking the parameter values above as part of the support set (see the right plot in Figure 1 for illustration). It may happen that for certain choices of k the support set is empty because the data do not constitute relative evidence at that level.

The choice of the support level is arbitrary, just as the choice of the confidence level from a confidence set is. One may, for example, report the support level $k = 1$ as it represents the tipping point at which the parameter values begin to be supported over the alternative. Conventions for Bayes factor evidence levels can also be used. For example, based on the conventions from Jeffreys (1961), a support set at level $k = 10$ includes the parameter values that receive 'strong' relative support from the data, while a $k = 1/10$ support set includes the parameter values that are at least not strongly contradicted.

2.2 The maximum evidence estimate

A natural point estimate for the unknown parameter θ based on the BFF is given by

$$\hat{\theta}_{\text{ME}} = \arg\max_{\theta_0 \in \Theta} \text{BF}_{01}(y; \theta_0),$$

and we call it the *maximum evidence estimate* (MEE), since it is the parameter value for which the Bayes factor indicates the most evidence over the alternative. The associated *evidence level*

$$k_{\text{ME}} = \text{BF}_{01}(y; \hat{\theta}_{\text{ME}}),$$

that is, the BFF evaluated at the MEE, quantifies the evidential value of the estimate $\hat{\theta}_{\text{ME}}$ over the alternative. Evidence levels close to $k_{\text{ME}} = 1$ indicate that the MEE receives little support over the alternative hypothesis H_1, whereas large evidence levels k_{ME} indicate that the MEE receives substantial support over the alternative hypothesis H_1. A useful summary of a BFF could hence be to report the MEE, its evidence level, and a support set, similar to how a P-value function may be summarized with a point estimate and confidence set.

To understand behaviour of the MEE with increasing sample size, we may look at an approximation of the Bayes factor. Suppose that the data $Y_{1:n} = \{Y_1, Y_2, \ldots, Y_n\}$ are independent and identically distributed and denote by $\hat{\psi}_0$ the maximizer of the log likelihood of the data under the null and by $(\hat{\theta}_1, \hat{\psi}_1)$ the maximizer under the alternative hypothesis. Denote by nV_0 and nV_1 the modal dispersion matrices (minus the inverse of the matrix of second-order partial derivatives of the log likelihood evaluated at the corresponding maximizer). Applying a Laplace approximation to the logarithm of

the BFF (O'Hagan and Forster, 2004, equation 7.27) gives then

$$\log \text{BF}_{01}(Y_{1:n}; \theta_0) = \log \frac{p(Y_{1:n} \mid \theta_0, \hat{\psi}_0)}{p(Y_{1:n} \mid \hat{\theta}_1, \hat{\psi}_1)} + \frac{\dim(\theta)}{2} \log \frac{n}{2\pi} + \log \frac{p(\hat{\psi}_0 \mid H_0)}{p(\hat{\theta}_1, \hat{\psi}_1 \mid H_1)} + \frac{1}{2} \log \frac{|V_0|}{|V_1|} + \mathcal{O}\left(n^{-1}\right). \tag{2}$$

To obtain the MEE, the log Bayes factor (2) needs to be maximized with respect to θ_0. It is clear that as θ_0 becomes more different from $\hat{\theta}_1$, the log normalized profile likelihood (first term) will decrease toward negative infinity, indicating evidence against the parameter value θ_0. On the other hand, when θ_0 is not too far from $\hat{\theta}_1$ the term will be about zero, so that an increasing sample size n (second term) increases the log BFF toward positive infinity, indicating evidence for θ_0. The relative accuracy of the priors (third term) and the relative dispersion (fourth term) lead to further adjustments of the BFF. For instance, when a parameter estimate is likely under the corresponding prior, this increases the evidence for corresponding hypothesis while a misspecified prior that is in conflict with the parameter estimates lowers the evidence for the corresponding hypothesis. In sum, finding the MEE corresponds to maximizing the profile likelihood that is adjusted based on the accuracy of the prior for the nuisance parameters and the modal dispersion.

2.3 Example: Normal mean

Assume we observe a single observation y assumed to be sampled (at least approximately) from a normal distribution $Y \mid \theta \sim \text{N}(\theta, \sigma^2)$. Assume that σ^2 is known and we want to conduct inferences regarding θ. This is a simple but frequently encountered scenario, for example, y could be an estimated regression coefficient from a generalized linear model or from a Cox proportional hazard model, and σ its estimated standard error. In the following we will consider an example from RECOVERY collaborative group (2022). This randomised controlled trial found a reduction in mortality of patients hospitalised with COVID-19 when treated with baricitinib compared to usual care (age-adjusted log hazard ratio $y = -0.14$ with standard error $\sigma = 0.064$ estimated with Cox regression). To obtain a Bayes factor for contrasting $H_0 \colon \theta = \theta_0$ against $H_1 \colon \theta \neq \theta_0$ we need to formulate a prior for θ under the alternative H_1. We will now discuss three choices with different characteristics shown in Table 1.

Perhaps the simplest choice is a prior that does not depend on the parameter value θ_0 of the null hypothesis, for example, a normal prior with mean m and variance v (left column in Table 1). The hyperparameters m and v may be specified based on external data or based on an alternative hypothesis of interest (e.g., the prior mean m could be set to a minimum clinically important treatment effect and v could be set to zero to obtain a point prior as typically used in a power analysis). For example, RECOVERY collaborative group (2022) reported a meta-analytic log hazard ratio and standard error based on eight previous trials, which could be used to set the prior mean and variance to $m = -0.56$ and variance $v = 0.12^2$, see Figure 2 for the resulting BFF (orange). In this case, the MEE is given by $\hat{\theta}_{\text{ME}} = y = -0.14$ with the support interval centered around it. Due to the apparent conflict between the observed data and the specified prior under the alternative, the $k = 1$ support interval spans an wide range from -0.35 to 0.08, indicating that very beneficial up to slightly harmful treatment effects are supported by the data over the alternative.

The formulae in Table 1 (left column) show that as the prior mean m becomes closer to the ob-

Table 1: Bayes factor function BF_{01}, maximum evidence estimate $\hat{\theta}_{\text{ME}}$, evidence value k_{ME}, and k support interval (SI) for a normal mean based on one observation y from $Y \mid \theta \sim \text{N}(\theta, \sigma^2)$ with known variance σ^2 and for three prior distributions for θ under the alternative H_1: A normal prior (left), a normal prior centered around the parameter value of the null hypothesis θ_0 (middle), a point prior shifted away from the parameter value of the null hypothesis θ_0 by $d > 0$ (right).

	Prior for θ under H_1		
	$\theta \sim \text{N}(m, v)$	$\theta \sim \text{N}(\theta_0, v)$	$\theta = \theta_0 + d$
BF_{01}	$\exp\left[-\frac{1}{2}\left\{\frac{(y-\theta_0)^2}{\sigma^2} - \frac{(y-m)^2}{\sigma^2+v}\right\}\right]\sqrt{1+\frac{v}{\sigma^2}}$	$\exp\left[-\frac{1}{2}\left\{\frac{(y-\theta_0)^2}{\sigma^2(1+\sigma^2/v)}\right\}\right]\sqrt{1+\frac{v}{\sigma^2}}$	$\exp\left\{\frac{2d(\theta_0-y)+d^2}{2\sigma^2}\right\}$
$\hat{\theta}_{\text{ME}}$	y	y	non-existent
k_{ME}	$\exp\left\{\frac{(y-m)^2}{2(\sigma^2+v)}\right\}\sqrt{1+\frac{v}{\sigma^2}}$	$\sqrt{1+\frac{v}{\sigma^2}}$	non-existent
k SI	$y \pm \sigma\sqrt{\log(1+\frac{v}{\sigma^2}) + \frac{(y-m)^2}{\sigma^2+v} - \log k^2}$	$y \pm \sigma\sqrt{\left\{\log(1+\frac{v}{\sigma^2}) - \log k^2\right\}(1+\frac{\sigma^2}{v})}$	$\left[y + \frac{\sigma^2 \log k}{d} - \frac{d}{2}, \infty\right]$

served data y, the evidence level k_{ME} decreases and the support interval becomes narrower. This is because an alternative closer to the data clearly has better predictive accuracy of the data than an alternative further away. Figure 2 illustrates this phenomenon with another prior distributions (one with mean at the observed log hazard ratio $y = -0.14$, the blue BFF), which is still centered at the observed log hazard ratio estimate but with far narrower $k = 1$ support interval from -0.22 to -0.06 than the orange BFF with the mean $m = -0.56$ based on the eight previous trials.

Another approach to formulating a prior distribution for θ under the alternative commonly suggested in 'objective' Bayes theories is to center the prior around the tested parameter value θ_0 (Jeffreys, 1961; Berger and Delampady, 1987; Kass and Wasserman, 1995). For example, one can specify a normal prior with mean at the null value θ_0 (middle column in Table 1). Thus, unlike the 'global' normal prior with fixed mean m, the resulting BFF varies both the null and the alternative. As a result, the interpretation of the BFF is different: For such a 'local' normal prior, the BFF quantifies the support of parameter values over alternative parameter values in a neighborhood around them. As for the global normal prior, the MEE based on the local normal prior is given by $\hat{\theta}_{\text{ME}} = y$ and support intervals are centered around it, but the associated, Bayes factor, evidence level and support interval are different. Figure 2 illustrates with the data from the RECOVERY trial that when the mean m of a global normal prior is too different from the observed data y (as in the case of the orange BFF, where the prior was specified based on the eight previous trials), the $k = 1$ support interval based on the local prior with the same variance is narrower. On the other hand, when the mean m of the global prior is equal to the data (blue BFF), the support interval based on the local prior is wider.

The last prior in the right most column of Table 1 represents a point prior shifted from the null value θ_0 by $d > 0$. The prior is again 'local' in the sense that it is different for each tested parameter value of the null hypothesis θ_0, and as such encodes an alternative hypothesis that the log hazard ratio is greater than the tested parameter value. However, this leads to an ever-increasing BFF, see Figure 2 for a numerical illustration. As a result, the MEE and its evidence level do not exist, while

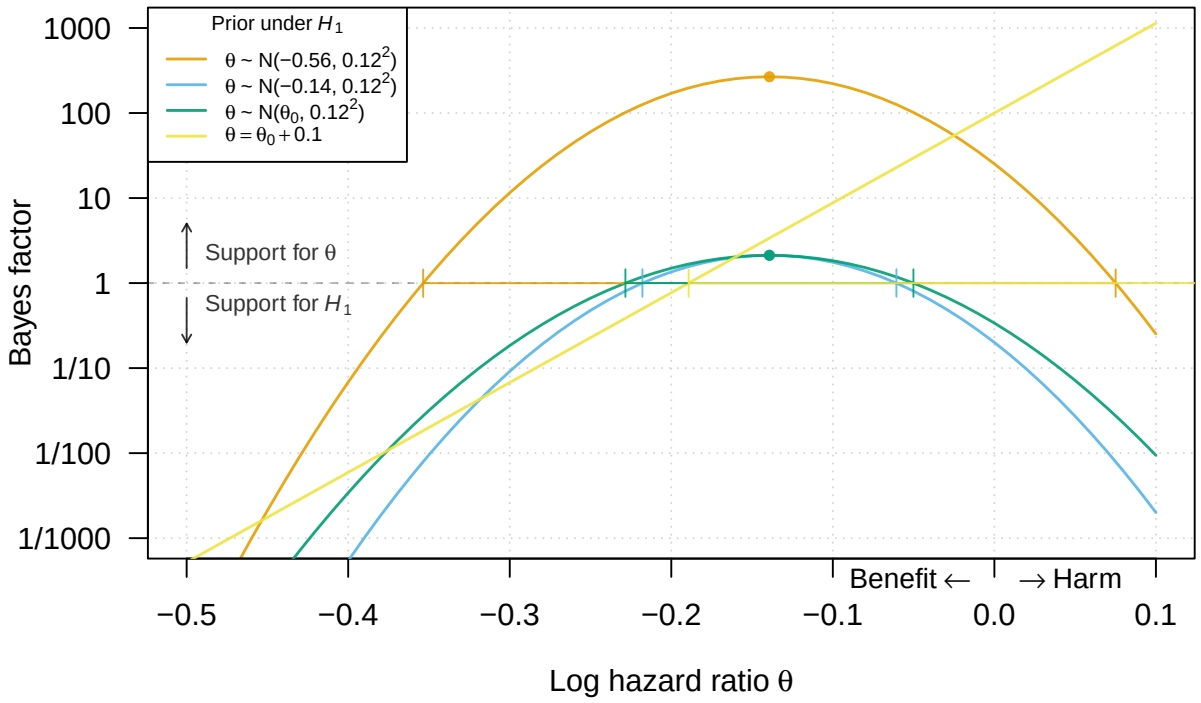

Figure 2: Bayes factor function, MEE, and $k = 1$ support interval for a log hazard ratio θ based on estimated log hazard ratio $y = -0.14$ with standard error $\sigma = 0.064$ from the RECOVERY trial (RECOVERY collaborative group, 2022) for different prior distributions for the θ under the alternative H_1. A normal likelihood $Y \mid \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ is assumed for the data.

the support interval still exists but its right limit extends to infinity. Although such a prior seems unrealistic, the example demonstrates that a poorly chosen prior can lead to pathological behavior of the resulting BFF.

2.4 Choice of the prior

As the previous example showed, the prior assigned to the parameters under the alternative has a substantial impact on BFF inference. This 'sensitivity' of Bayes factors to prior distributions enables data analysts to accurately quantify the support of parameter values over informative alternative hypotheses when they are available, but poses a challenge in their absence (Kass and Raftery, 1995). Various approaches have been proposed to deal with this issue, for example, 'default' or 'objective' prior distributions (Bayarri et al., 2012; Consonni et al., 2018), reverse-Bayes analysis (Held et al., 2022), prior elicitation (O'Hagan, 2019), or sensitivity analysis (Franck and Gramacy, 2019), all with advantages and disadvantages. Here we will not reiterate general considerations on prior specification for Bayes factors (see e.g., Section 5 in Kass and Raftery, 1995) but focus on specific considerations related to BFFs.

As in other Bayes factor applications, BFFs are only unambiguously defined if priors for focus parameters are proper under the alternative H_1 (i.e., integrate to one), whereas priors for nuisance parameters may be improper as long as the same prior is assigned under both the null H_0 and the alternative H_1 so that arbitrary constants cancel out. A general distinction can be made between

global priors, which do not depend on the value of θ_0 under the null hypothesis and *local* priors, which do. In the latter case, the interpretation of the BFF is more intricate, since for each parameter value the BFF quantifies the support over a different alternative. For a more natural interpretation, global priors may hence be preferred over local priors. At the same time, local priors correspond to the typical use of 'default' Bayes factors, which is to center the prior around θ_0, and as such may be preferred in the same situations where default Bayes factors would be used.

Finally, it is usually advisable to report sensitivity analyses for plausible ranges of priors, to assess the robustness of the conclusions. A convenient visual sensitivity analysis is, for example, to plot different BFFs resulting from different prior specifications, as shown in Figure 2. One can go a step further and use a 'reverse-Bayes' approach (Good, 1950; Held et al., 2022), which involves systematically determining the prior that represents the tipping point and changes the conclusions of the analysis. Data analysts can then reason about whether or not such a prior is plausible in the light of external knowledge and data.

2.5 Sequential analysis

An attractive property of Bayesian inference is that it provides a coherent way to analyze data that come in batches. That is, the same posterior distribution is obtained regardless of whether all data are analyzed at once, or whether the posterior distribution based on one batch is used as the prior for the other.

If we have two batches y_1 and y_2, the BFF based on both batches is

$$\text{BF}_{01}(y_1, y_2; \theta_0) = \text{BF}_{01}(y_1; \theta_0) \times \text{BF}_{01}(y_2 \mid y_1; \theta_0)$$

where

$$\text{BF}_{01}(y_2 \mid y_1; \theta_0) = \frac{\int_\Psi p(y_2 \mid \theta_0, \psi)\, p(\psi \mid y_1, H_0)\, \mathrm{d}\psi}{\int_\Theta \int_\Psi p(y_2 \mid \theta, \psi)\, p(\theta, \psi \mid y_1, H_1)\, \mathrm{d}\psi\, \mathrm{d}\theta}$$

is the partial Bayes factor obtained from using the posterior distributions $p(\psi \mid y_1, H_0)$ and $p(\theta, \psi \mid y_1, H_1)$ based on the first batch y_1 to compute the Bayes factor based on the second batch y_2 (O'Hagan and Forster, 2004, p.186). This result generalizes to more than two batches by

$$\text{BF}_{01}(y_1, y_2, \ldots, y_n; \theta_0) = \text{BF}_{01}(y_1; \theta_0) \times \prod_{i=2}^{n} \text{BF}_{01}(y_i \mid y_1, y_2, \ldots, y_{i-1}; \theta_0),$$

that is, a BFF based on all the available data can be obtained by updating the BFF based on the previous batches by the partial Bayes factor. Thus, like ordinary Bayesian inference with posterior distributions, BFF inference is sequentially coherent.

2.6 Asymptotic behaviour of the Bayes factor function

It is of interest to understand the asymptotic behaviour of the BFF, that is, how does the BFF (and quantities derived from it) behave as more data are generated under a certain 'true' hypothesis? It is well-known that Bayes factors are consistent in the sense that when the data are generated under one

of the contrasted hypotheses, the Bayes factor tends to overwhelmingly favour that hypothesis over the alternative as more data are generated, i.e., go to zero or infinity, depending on the orientation of the Bayes factor (see e.g., Dawid, 2011). Since the BFF is nothing else than the Bayes factor evaluated for various null hypotheses, this consistency property carries over to the BFF. That is, as more data are generated from the model with true parameter θ_*, the BFF at $\theta_0 = \theta_*$ will go to infinity, while the BFF at $\theta_0 \neq \theta_*$ will go to zero.

As a concrete example where the distribution of the BFF can be derived in closed-form, consider again inference about a normal mean based on data $Y \mid \theta \sim \mathrm{N}(\theta, \kappa^2/n)$, where κ^2 denotes a unit-variance and n the sample size. The logarithm of the BFF based on a normal prior $\theta \mid H_1 \sim \mathrm{N}(m, v)$ can then be written as

$$\log \mathrm{BF}_{01}(Y; \theta_0) = \frac{1}{2}\left[\log\left(1 + \frac{n\,v}{\kappa^2}\right) + \frac{(\theta_0 - m)^2}{v} - \left\{Y - \frac{(\theta_0 - m)\kappa^2}{n\,v} - \theta_0\right\}^2 \frac{v\,n}{\kappa^2(v + \kappa^2/n)}\right]. \quad (3)$$

Hence, when the data are generated from $Y \mid \theta_* \sim \mathrm{N}(\theta_*, \kappa^2/n)$ with true mean θ_*, we have that

$$\left\{Y - \frac{(\theta_0 - m)\kappa^2}{n\,v} - \theta_0\right\}^2 \frac{n}{\kappa^2} \sim \chi^2_{1,\lambda}$$

with non-centrality parameter $\lambda = n\left\{\theta_* - \frac{(\theta_0 - m)\kappa^2}{n\,v} - \theta_0\right\}^2 / \kappa^2$. Thus, by rearranging terms in (3), we can compute the probability that the Bayes factor is below some threshold γ by

$$\Pr(\mathrm{BF}_{01}(Y; \theta_0) \leq \gamma \mid \theta_*) = 1 - \Pr(\chi^2_{1,\lambda} \leq X)$$

with

$$X = \left\{\log\left(1 + \frac{n\,v}{\kappa^2}\right) + \frac{(\theta_0 - m)^2}{v} - 2\log\gamma\right\}\left(1 + \frac{\kappa^2}{v\,n}\right).$$

Figure 3 shows the distribution of the BFF for different sample sizes, a true mean of $\theta_* = 0$, a unit-variance of $\kappa^2 = 4$, and with a local normal prior with the same unit variance (a unit-information prior, see Kass and Wasserman, 1995) specified under the alternative. We see that as the sample size increases, the distribution of the BFF at the true mean shifts toward larger values, indicating more evidence for the true mean, as it should. On the other hand, the further away the BFF is evaluated from the true mean, the more its distribution shifts toward smaller values, indicating increasing evidence for the alternative, as it should.

3 Connection to other inference frameworks

We will now explore connections of BFF inference to other inference frameworks.

3.1 Maximum integrated likelihood

In typical situation where a division of $p(y \mid H_0)$ by $p(y \mid H_1)$ does not change the maximizer of $p(y \mid H_0)$, the MEE can be obtained by maximizing the marginal likelihood $p(y \mid H_0)$ without ref-

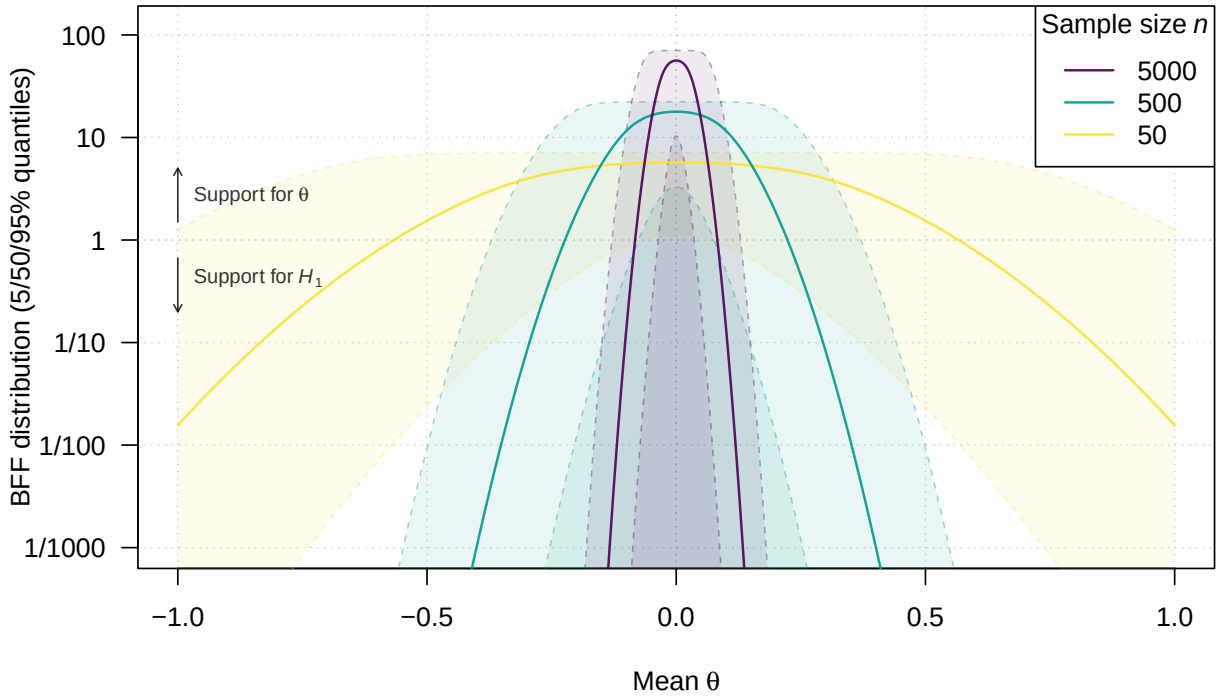

Figure 3: Distribution of the BFF for different sample sizes. A data model $Y \mid \theta \sim \mathrm{N}(\theta, \kappa^2/n)$ is assumed and data are generated from a true mean $\theta_* = 0$ and unit-variance $\kappa^2 = 4$. The BFF is based on a local normal prior $\theta \mid H_1 \sim \mathrm{N}(\theta_0, v = 4)$ assigned to θ under the alternative.

erence to an alternative H_1. This is, for instance, the case when a global prior (a prior that does not depend on θ_0) is assigned to θ under the alternative, or also in the case of the local normal prior that is centered around θ_0 from the previous example. The MEE is then equivalent to the maximizer of the *integrated likelihood*

$$\hat{\theta}_{\text{MIL}} = \arg\max_{\theta \in \Theta} \int_{\Psi} p(y \mid \theta, \psi)\, p(\psi \mid H_0)\, \mathrm{d}\psi,$$

based on prior $p(\psi \mid H_0)$ assigned to the nuisance parameters (see e.g., Kalbfleisch and Sprott, 1970; Basu, 1977; Berger et al., 1999; Royall, 1997; Severini, 2007). When there are no nuisance parameters, the MEE reduces to the ordinary maximum likelihood estimate.

To consider a concrete example, assume a sample of n normal random variables $Y_1, \ldots, Y_n \mid \theta, \sigma^2 \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$. Suppose that σ^2 is the focus and μ the nuisance parameter, and that an improper uniform prior $p(\mu \mid H_0) = 1$ is assigned to μ. The intergrated likelihood of an observed sample $y_1, \ldots, y_n$ is then

$$p(y_1, \ldots, y_n \mid \sigma^2) = (2\pi\sigma^2)^{-(n-1)/2}\, n^{-1/2} \exp\left\{ \frac{-\sum_{i=1}^{n}(y_i - \bar{y})^2}{2\sigma^2} \right\}$$

and maximizing it leads to the sample variance (REML) estimate of the variance

$$\hat{\sigma}^2_{\text{MIL}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

The same MEE is obtained when the prior $p(\mu, \sigma^2 \mid H_1)$ does not depend on the value of the variance under H_0 as the denominator of the Bayes factor is simply a multiplicative factor that does not change its maximum. This shows that REML estimates can be motivated from a Bayesian evidence perspective which complements the well-established connections between REML estimation and marginal posterior estimation based on flat priors for the nuisance parameters (Harville, 1974; Laird and Ware, 1982). It is reassuring that different methods produce the same estimate in these situations. However, the important difference between these methods is the motivation and interpretation of the resulting estimate – the MEE represents a natural estimate for θ because it is the parameter value for which the data provide the most evidence over an alternative hypothesis, while the MLE is defined without reference to alternatives.

3.2 Likelihoodist inference

The likelihoodist school of statistical inference (Barnard, 1949; Edwards, 1971; Royall, 1997; Blume, 2002) rejects the use of prior distributions to formulate alternatives or to eliminate nuisance parameters, but it also shares features with the BFF paradigm. That is, if point priors are assigned to the parameters, the Bayes factor reduces to a likelihood ratio which is the evidence measure used by likelihoodists. For this reason, BFF inferences correspond to likelihoodist inferences if the Bayesian and likelihoodist agree on the used point priors.

However, there is disagreement when it comes to the use of support sets. When there are no nuisance parameters, likelihoodists define their support sets based on the relative likelihood

$$L(\theta) = \frac{p(y \mid \theta)}{p(y \mid \hat{\theta}_{\text{ML}})}. \tag{4}$$

For example, Royall (1997) recommended reporting the set of parameter values with relative likelihood greater than $k = 1/8$ (at most 'strong' evidence against them) or $k = 1/32$ (at most 'quite strong' evidence against them). From a Bayesian perspective, using the observed MLE as a prior under the alternative seems to hardly represent genuine prior knowledge or an alternative theory, but rather a cherry-picked alternative that gives to the most biased assessment of support for the alternative (Berger and Sellke, 1987).

3.3 Frequentist inference

The relative likelihood (4) serves as an important basis for frequentist statistics since under the null hypothesis $-2 \log L(\theta_0)$ has an asymptotic chi-squared distribution with $\dim(\theta)$ degrees of freedom. Frequentists thus also use relative likelihoods but merely as a test statistic.

Another connection between frequentist and BFF inference is given by the 'universal bound' (Kerridge, 1963; Robbins, 1970; Royall, 1997), which bounds the frequentist probability of obtaining misleading Bayesian evidence. That is, for $0 < k < 1$ the probability of obtaining a Bayes factor against H_0 less than k is at most k for any prior under the alternative

$$\Pr\{\text{BF}_{01}(y; H_0) \leq k \mid H_0\} \leq k.$$

If there are nuisance parameters, the bound holds only marginalized over the prior of the nuisance parameters. For the bound to hold in a strict sense (i.e., for every possible value of the nuisance parameter), special priors must be assigned to them (Hendriksen et al., 2021; Grünwald et al., 2024).

The universal bound can thus be used to transform BFFs into conservative P-values and confidence sets, e.g., a $k = 1/20$ support set obtained from a BFF corresponds to a 95% conservative confidence set and $p = \max\{\text{BF}_{01}, 1\}$ corresponds to a conservative P-value. Remarkably, the bound holds without adjustment even when the data collection is continuously monitored and stopped as soon as evidence against H_0 is found (Robbins, 1970). However, it is important to note that P-values and confidence sets obtained in this way are usually much more conservative than ordinary ones which are calibrated to have exact type I error rate and coverage, respectively. Finally, if the data model is misspecified, the bound is obviously invalid.

3.4 Bayesian inference

The BFF can, under certain conditions, be transformed into a Bayesian posterior distribution. Specifically, assuming a 'global' prior under the alternative, a prior $p(\theta \mid H_1)$ which does not depend on the parameter under the null θ_0, and also that the priors for the nuisance parameters satisfy $p(\psi \mid H_0) = p(\psi \mid \theta = \theta_0, H_1)$, we have the well-known Savage-Dickey density ratio (Dickey, 1971; Verdinelli and Wasserman, 1995; Wagenmakers et al., 2010) result

$$p(\theta \mid y, H_1) = \underbrace{\frac{p(y \mid \theta, H_1)}{p(y \mid H_1)}}_{=\text{BF}_{01}(y;\theta)} \times p(\theta \mid H_1). \tag{5}$$

That is, the posterior can be obtained by multiplying the BFF with the prior. It is, however, important to emphasize that BFFs based on priors under the alternative that depend on the null (e.g., commonly used 'local' normal or Cauchy priors that are centered around θ_0) cannot be transformed to a genuine posterior distribution in this way, but multiplication with the prior will result in a different posterior for every θ.

The relative belief inference framework is perhaps the closest to the one presented here. However, as with the posterior distribution there is only a correspondence between relative belief ratio inferences and BFF inference if the prior for θ is the same for every θ_0, and it does not hold anymore when, for example, local priors centered around the value of the null θ_0 are used, or when the prior for the nuisance parameters under the null differs from the prior under the alternative.

Since, under certain regularity conditions, the posterior is asymptotically normally distributed around the maximum likelihood estimate with a covariance matrix equal to the observed information matrix (Bernardo and Smith, 2000, chapter 5.3), we can conclude that whenever the BFF has the Savage-Dickey density ratio representation (5), asymptotically the BFF is the asymptotic posterior normal density divided by the prior density, both evaluated at θ_0. Hence, in contrast to the posterior distribution, the influence of the prior on the BFF does not 'wash' out as more data are collected.

The link (5) also provides a convenient way to compute BFFs when it applies: One of the many programs for computing Bayesian posterior distributions, such as Stan (Carpenter et al., 2017) or INLA (Rue et al., 2009), can be used to compute a posterior density, which can then be divided by

the prior density to obtain a BFF. The caveat is again that this only works for global priors under the alternative and with the same prior assigned to the nuisance parameters under the null and alternative.

The relationship between the posterior and the BFF also exposes its connection to another Bayesian inference quantity – the *relative belief ratio*

$$\text{RB}(\theta \mid H_1) = \underbrace{\frac{p(\theta \mid y, H_1)}{p(\theta \mid H_1)}}_{=\text{BF}_{01}(y;\theta)}, \tag{6}$$

see e.g., Evans (2015). This quantity is the updating factor of the prior to the posterior density/probability mass function, and is related to the Bayes factor via the aforementioned mentioned Savage-Dickey density ratio. An estimation and testing framework centred on the relative belief ratio was developed by Evans (1997). The parameter value that maximizes the relative belief ratio was termed the *least relative surprisal estimate*, later also referred to as *maximum relative belief estimate* (Evans, 2015). Clearly this estimate equivalent to the MEE whenever the BFF and relative belief ratio coincide. Evans (1997) also defined a γ *relative surprise region*, which is the set of parameter values with γ posterior probability and with highest relative belief ratios among all such sets. Similarly, Shalloway (2014) defined an *evidentiary credible region* which is equivalent to the relative surprise region, but motivated by information theory. While both are closely related to the support set via the Savage-Dickey density ratio, they differ from the support set in that they are defined by posterior probabilities and not by the minimum evidence that the parameter values that they contain receive (Wagenmakers et al., 2020). Thus, a relative surprise region may contain parameter values that are not supported by the data. For this reason, Evans (2015) defined yet another type of region, a q *plausible region* which contains parameter values with a relative belief ratio of at least q and as such coincides with the k support set whenever the Savage-Dickey density representation applies to the BFF.

4 Applications

We will now illustrate application of the BFF inference framework on several real-world examples.

4.1 Binomial proportion

Bartoš et al. (2023) conducted a study to test the hypothesis that fair coins tend to land on the same side as they started slightly more often (with a probability of about 0.51). This hypothesis was formulated by Diaconis et al. (2007) based on a physical model of coin flipping. During the course of the study, 48 participants contributed to the collection of $n = 350'757$ coin flips among which $y = 178'078$ landed on the same side as they started.

We will now assume a binomial data model $Y \mid \theta \sim \text{Bin}(n, \theta)$ and conduct inferences regarding the unknown probability θ. In their pre-registered analysis, Bartoš et al. (2023) specified a truncated beta prior for the probability θ under the alternative $(\theta \mid H_1 \sim \text{Beta}(a, b)_{[l,u]})$. Based on this prior, the

Bayes factor for testing $H_0 \colon \theta = \theta_0$ against $H_1 \colon \theta \neq \theta_0$ is

$$\text{BF}_{01}(y; \theta_o) = \frac{\theta_0^y \, (1 - \theta_0)^{n-y}}{\text{B}(a + y, b + n - y) \,/\, \text{B}(a, b)} \times \frac{I_u(a, b) - I_l(a, b)}{I_u(a + y, b + n - y) - I_l(a + y, b + n - y)}$$

with the beta function $\text{B}(a, b) = \int_0^1 t^{a-1} \, (1 - t)^{b-1} \, \text{d}t$ and the incomplete regularized beta function $I_x(a, b) = \{\int_0^x t^{a-1} \, (1 - t)^{b-1} \, \text{d}t\} \,/\, \text{B}(a, b)$. Specifically, Bartoš et al. (2023) assigned the hyperparameters $a = 5100, b = 4900, l = 0.5, u = 1$ to instantiate an alternative hypothesis that closely aligns with the theoretical prediction from Diaconis et al. (2007) of a 0.51 probability with slight uncertainty around it.

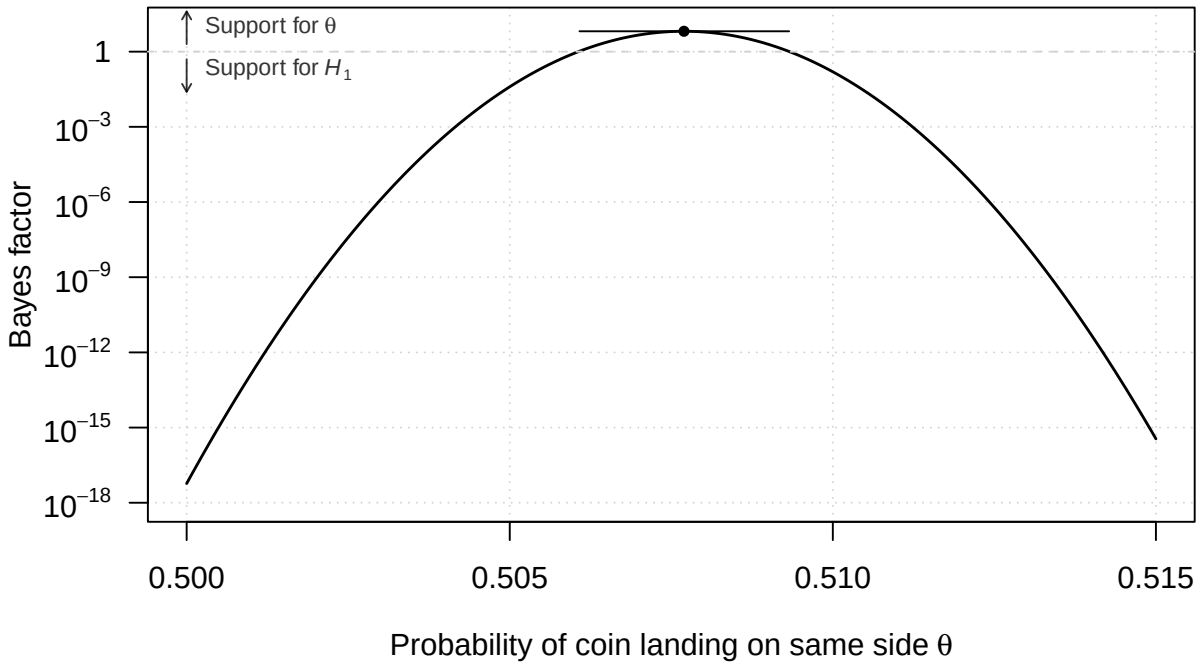

Figure 4: Bayes factor function analysis of data from Bartoš et al. (2023): from $n = 350'757$ coin flips, $y = 178'078$ landed on the same side as they started. A beta prior tightly concentrated around the theoretically predicted probability of 51% is assigned to the probability under the alternative $(\theta \mid H_1 \sim \text{Beta}(5100, 4900)_{[0.5,1]})$.

Figure 4 shows the resulting BFF for a range of probabilities from 0.5 to 0.515. Looking at the BFF evaluated at $\theta = 0.5$, we can see the result reported by Bartoš et al. (2023): There is extreme evidence ($\text{BF}_{01} = 1/(1.71 \times 10^{17})$) against $\theta = 0.5$ and in favour of the alternative concentrated around $\theta = 0.51$. This result hence provides decisive evidence against the hypothesis that coins tend to land on the same side with equal probability. However, the BFF framework permits further insights. For example, we can see that all probability values up to about 0.504 and all values larger than 0.512 are decisively refuted by the data, each having an associated Bayes factor below 10^{-3}. Furthermore, the $k = 1$ support interval from 0.506 to 0.509 shows the probability values that are better supported by the data than the specified alternative, which excludes the theoretically predicted $\theta = 0.51$. The MEE at $\hat{\theta}_{\text{ME}} = 0.508$ is the best supported value, with $k_{\text{ME}} = 6.51$ indicating substantial

evidence over the alternative concentrated around 0.51.

4.2 Meta-analysis

The previous analysis assumed that coin flips were independent among participants and trials. The top left plot in Figure 5 shows that this assumption seems violated as the estimated probabilities that a coin lands on the same side for each of the 48 study participants are clearly heterogeneous. This suggests that the analysis should be modified to account for heterogeneity. In the following, we will therefore synthesize these estimates while accounting for heterogeneity with a meta-analysis, as Bartoš et al. (2023) did.

Suppose we have $i = 1, \ldots, n$ estimates y_i with (assumed to be known) standard errors σ_i. The estimates are assumed to be normally distributed around a subject specific parameter θ_i, i.e.,

$$\begin{aligned} y_i \mid \theta_i, \sigma_i^2 &\sim \mathrm{N}(\theta_i, \sigma_i^2) \\ \theta_i \mid \theta, \tau^2 &\sim \mathrm{N}(\theta, \tau^2). \end{aligned}$$

Marginalized over the study-specific parameters, the distribution of an estimate is then

$$y_i \mid \theta, \tau, \sigma_i^2 \sim \mathrm{N}(\theta, \sigma_i^2 + \tau^2).$$

There are two unknown parameters, θ and τ. The mean θ quantifies the average true parameter across units (participants, studies, etc.), while the heterogeneity standard deviation τ quantifies the heterogeneity of these true parameters. The Bayes factor for testing $H_0 \colon \theta = \theta_0, \tau = \tau_0$ against $H_1 \colon \theta \neq \theta_0, \tau \neq \tau_0$ is then given by

$$\mathrm{BF}_{01}(y_1, \ldots, y_n; \theta_0, \tau_0) = \frac{\prod_i^n \mathrm{N}(y_i \mid \theta_0, \tau_0^2)}{\int_0^\infty \int_{-\infty}^{+\infty} \prod_i^n \mathrm{N}(y_i \mid \theta, \tau^2)\, p(\theta, \tau \mid H_1)\, \mathrm{d}\theta\, \mathrm{d}\tau}$$

with $\mathrm{N}(x \mid m, v)$ denoting the normal density with mean m and variance v evaluated at x.

As in the previous analysis we assigned a $\theta \mid H_1 \sim \mathrm{Beta}(5100, 4900)_{[0.5,1]}$ prior to the average probability θ under the alternative H_1. In addition, we assigned a half-normal prior $p(\tau \mid H_1) = \sqrt{2/\pi}\, \exp\{-\tau^2/(2\, s^2)\}/s$ to the heterogeneity standard deviation τ, and assumed it to be independent of θ. Half-normal priors are commonly used in meta-analysis due to their simplicity and desirable properties such as nearly uniform behavior around zero $\tau = 0$ (see e.g., Röver et al., 2021). We choose a scale $s = 0.02$ because the resulting prior gives a 95.4% probability to τ values smaller than 0.04, thus encoding the possibility of no heterogeneity (all participant probabilities are the same when $\tau = 0$) up to small amounts of heterogeneity (the true participant probabilities differ by a few percentage points). Several BFFs for priors with smaller or larger scale parameters are also shown in Figure 5 as sensitivity analyses.

The top-right plot in Figure 5 shows the BFF in a two-dimensional surface when both parameters are considered as focus parameters. In contrast to the analysis that ignored between-participant heterogeneity, we see that the MEE for the average probability ($\hat{\theta}_{\mathrm{ME}} = 0.51$) is now consistent with the theoretical prediction of Diaconis et al. (2007). In addition, the MEE for the heterogeneity standard

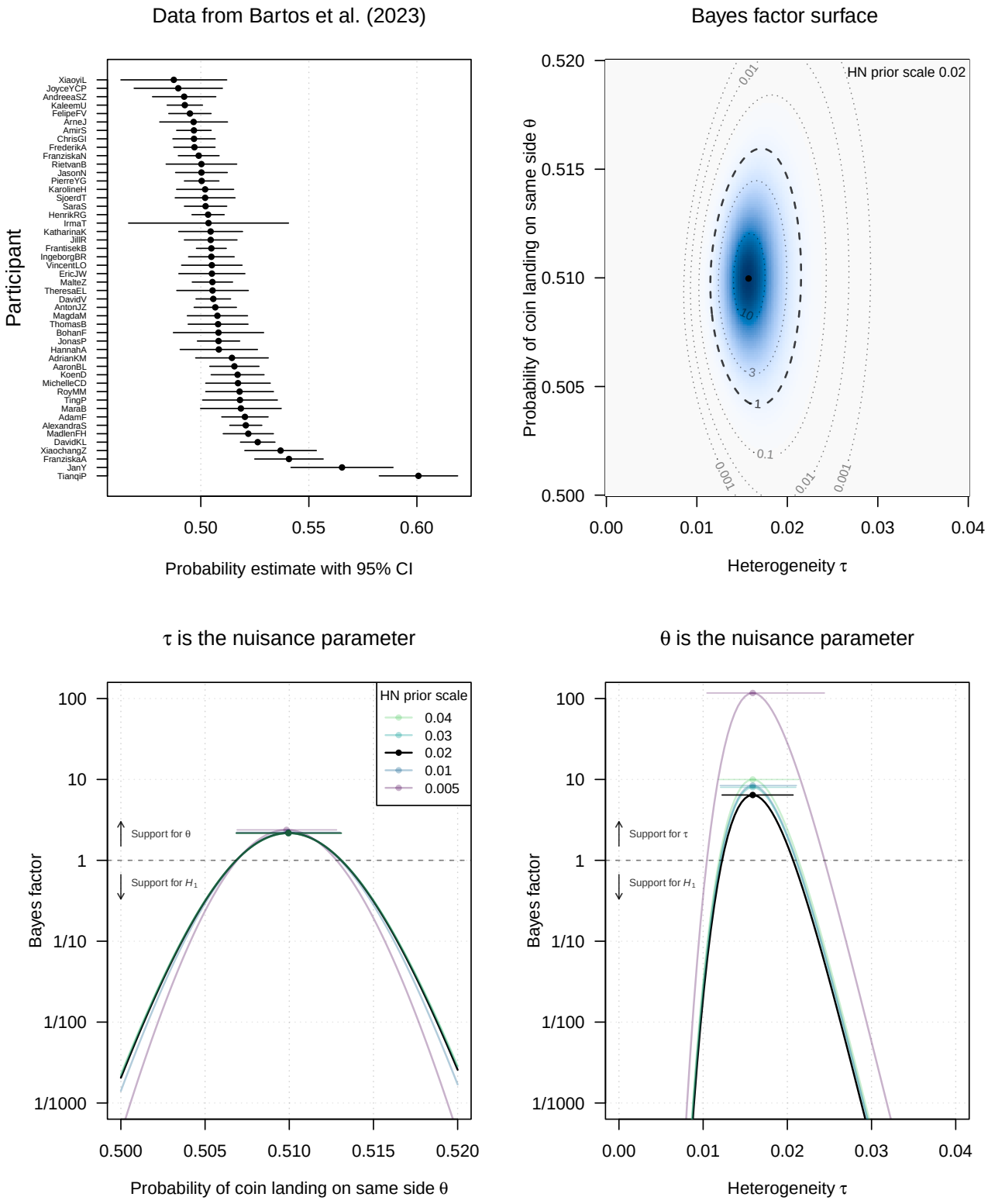

Figure 5: Bayes factor analysis of coin flipping experiments from Bartoš et al. (2023), taking into account between-participant heterogeneity. The product of a truncated beta prior ($\theta \mid H_1 \sim \text{Beta}(5100, 4900)_{[0.5,1]}$) for θ and a half-normal prior with scale 0.02 for τ are assigned under the alternative H_1. The same priors are assumed when the parameters are nuisance parameters under H_0 (bottom plots). The bottom plots also show the BFF for other scale parameters of the half-normal prior.

deviation ($\hat{\tau}_{\text{ME}} = 0.016$) suggests small but non-negligible heterogeneity. This MEE receives strong support over the alternative ($k_{\text{ME}} = 14$). The relatively concentrated $k = 1$ support region indicates that probabilities from around 0.505 to 0.515 along with heterogeneity standard deviations from 0.012 to 0.021 are supported by the data over the alternative. Finally, the BFF shows that probabilities of $\theta = 0.5$ and no heterogeneity $\tau = 0$ are clearly refuted by the data over the alternative ($\log \text{BF}_{01} = -1.81 \times 10^5$).

The two bottom plots in Figure 5 show BFFs when either τ or θ is considered as nuisance parameter. In both cases, the same prior as for the alternative H_1 was assigned to the corresponding nuisance parameter under H_0. In addition, BFFs for other choices of the scale parameter of the half-normal prior were computed to assess the sensitivity of the results to this choice. We see that the two MEEs ($\hat{\theta}_{\text{ME}} = 0.51$ and $\hat{\tau}_{\text{ME}} = 0.016$) align with the joint MEEs, but their evidence values ($k_{\text{ME}} = 2.2$ and $k_{\text{ME}} = 6.4$, respectively) indicate less support over the alternative than for the joint one. Finally, looking at the colored BFFs obtained by changing the scale parameter of the half-normal prior assigned to τ, we see that the scale has little effect on inferences about the probability θ, but a more pronounced effect on inferences about τ. Increasing the scale of the prior does not seem to change the BFF too much, while decreasing the scale to a value of $s = 0.005$ dramatically increases the height of the BFF, increasing the support of the MEE and surrounding values over the alternative. This seems reasonable, since the data show clear signs of heterogeneity, while a prior with such a small scale would predict almost none.

4.3 Replication studies

In a replication study, researchers repeat an original study as closely as possible in order to assess whether consistent results can be obtained (National Academies of Sciences, Engineering, and Medicine, 2019). Various types of Bayes factor approaches have been proposed to quantify the degree to which a replication study has replicated an original study (Verhagen and Wagenmakers, 2014; Ly et al., 2018; Harms, 2019; Pawel and Held, 2022; Pawel et al., 2023a). A common idea is that the posterior distribution of the unknown parameters based on the data from the original study is used as the prior distribution in the analysis of the replication data. If the replication data support this prior distribution, this suggests replication success. We will now show how this idea translates to analyzing replication studies with BFFs.

Suppose that original and replication study provide an effect estimate y_o and y_r with standard error σ_o and σ_r, respectively. Each is supposed to be normally distributed around the underlying effect size θ with (assumed to be known) variance equal to its squared standard error, i.e., $y_i \mid \theta \sim \text{N}(\theta, \sigma_i^2)$ for $i \in \{o, r\}$. A 'replication BFF' may then be obtained by contrasting the null hypothesis $H_0 \colon \theta = \theta_0$ to the alternative $H_1 \colon \theta \sim \text{N}(y_o, \sigma_o^2)$, where the prior under the alternative is the posterior distribution of θ based on the original data and a flat prior for θ (Verhagen and Wagenmakers, 2014). This leads to the following BFF

$$\text{BF}_{01}(y_r; \theta_0) = \sqrt{1 + \frac{\sigma_o^2}{\sigma_r^2}} \exp\left[-\frac{1}{2}\left\{\frac{(y_r - \theta_0)^2}{\sigma_r^2} - \frac{(y_r - y_o)^2}{\sigma_r^2 + \sigma_o^2}\right\}\right]$$

with MEE at the replication effect estimate $\hat{\theta}_{\text{ME}} = y_r$, evidence value

$$k_{\text{ME}} = \sqrt{1 + \frac{\sigma_o^2}{\sigma_r^2}} \exp\left\{\frac{-(y_r - y_o)^2}{2(\sigma_o^2 + \sigma_r^2)}\right\},$$

and k support interval

$$y_r \pm \sigma_r \sqrt{\log\left(1 + \frac{\sigma_o^2}{\sigma_r^2}\right) + \frac{(y_r - y_0)^2}{\sigma_r^2 + \sigma_o^2} - \log k^2}.$$

We will now reanalyze data from three replication studies that were part of the large-scale replication project in the social-behavioral sciences (Protzko et al., 2023). The original experiment termed "Label" found the following central result:

> "*When a researcher uses a label to describe people who hold a certain opinion, he or she is interpreted as disagreeing with that opinion when a negative label is used and agreeing with that opinion when a positive label is used.*" Protzko et al. (2023, p. 2)

which was based on an estimated standardized mean difference $y_o = 0.205$ with standard error $\sigma_o = 0.051$. The replication studies conducted in three other labs found a smaller, a similar, and a much larger effect estimate ($y_{ri} \in \{0.09, 0.205, 0.435\}$ with standard errors $\sigma_{ri} \in \{0.052, 0.057, 0.044\}$, respectively). The top plot in Figure 6 shows the associated BFFs, MEEs, and $k = 1$ support intervals. We see that, the BFFs peak at the corresponding replication estimate, but the height of these peaks differs between replications. The replication from lab 2 produced an effect estimate identical to the original one, so there is little support of its MEE over the alternative based on the original study as the estimates from both studies are in close agreement. In contrast, the MEEs from lab 1 and lab 3 receive substantial and very strong support over the alternative because they are either smaller or larger. In turn, their $k = 1$ support intervals are much wider compared to the narrow support interval from lab 2. Finally, we can also see that the BFF from lab 1 indicates absence of evidence for or against no effect ($\text{BF}_{01} \approx 1$ at $\theta = 0$), whereas the BFFs from labs 2 and 3 indicate strong and decisive evidence against no effect up to very small effects of around $\theta = 0.1$.

The bottom plot in Figure 6 illustrates the posterior distributions, conveniently obtained by multiplying the BFF by the prior distribution based on the original data. We can see that these posteriors represent a synthesis of the original and replication studies. For example, the posterior based on the replication from lab 3 is centered around 0.34 with 95% credible interval from 0.27 to 0.4. Clearly, this interval excludes both the original ($y_o = 0.205$) and the replication effect estimates ($y_{r3} = 0.435$), leading to an opposite conclusion from the BFF, which indicates decisive evidence for the MEE at y_r.

4.4 Logistic regression

To illustrate a computationally more involved application of BFFs, we consider the study from Neutra et al. (1978), previously reanalyzed by Greenland (2007); Sullivan and Greenland (2012). This epidemiological study investigated the association between 14 exposure variables and the neonatal death. Among the 2992 births only 17 neonatal deaths occurred. This leads to challenges in conducting inferences with so many exposure variables, as we will see in the following.

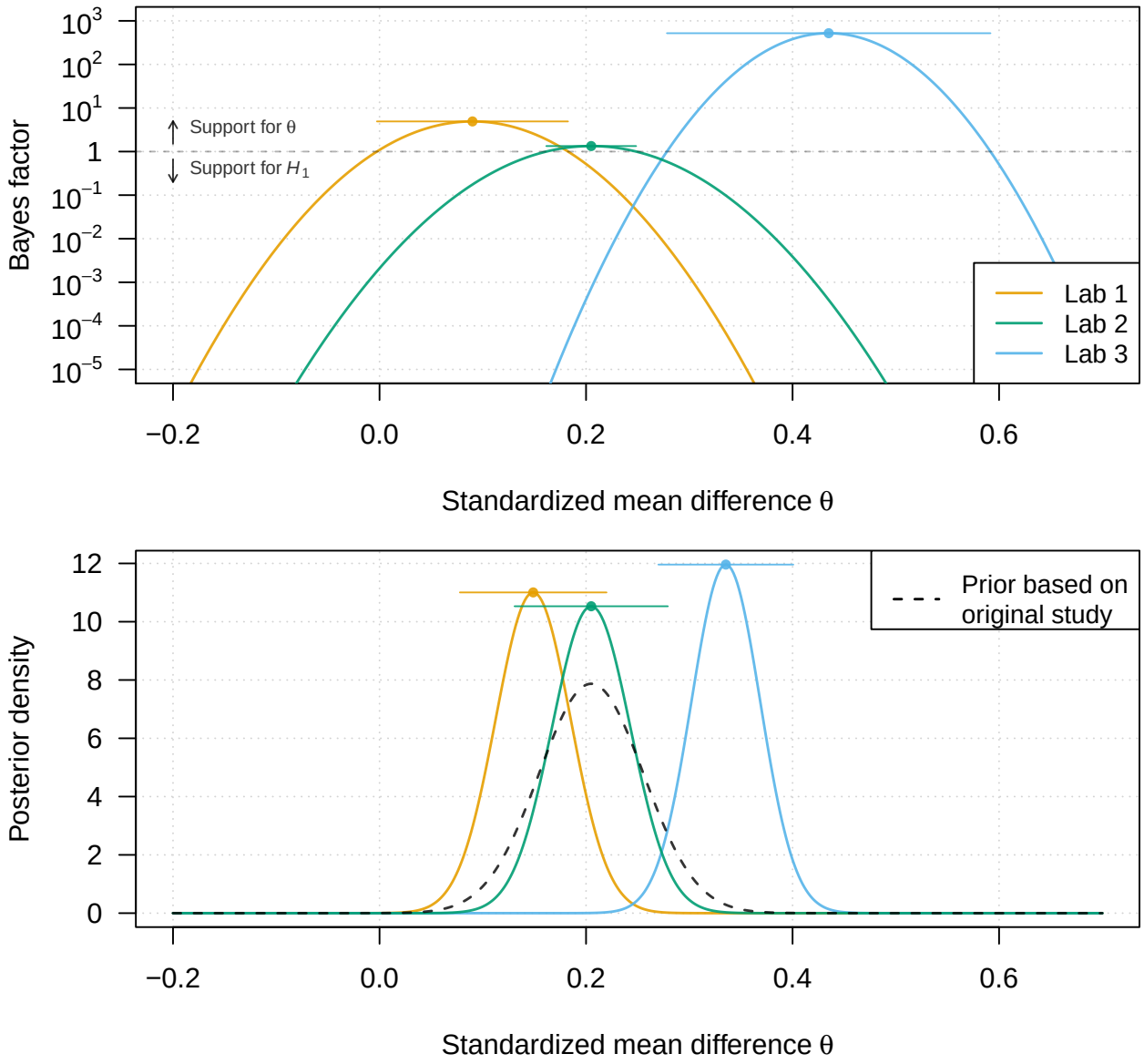

Figure 6: Bayes factor functions with maximum evidence estimates and $k = 1$ support intervals (top) and posterior distribution with posterior modes and 95% highest posterior density credible intervals (bottom) for the three replication studies from the "Labels" experiment (Protzko et al., 2023). The original study found an estimated standardized mean difference of $y_o = 0.205$ with standard error $\sigma_o = 0.051$ which is used to formulate the prior distribution under the alternative $\theta \mid H_1 \sim \mathrm{N}(y_o, \sigma_o^2)$. The replication effect estimates were $y_{ri} \in \{0.09, 0.205, 0.435\}$ with standard errors $\sigma_{ri} \in \{0.052, 0.057, 0.044\}$, respectively, and are assumed to be normally distributed $y_{ri} \mid \theta \sim \mathrm{N}(\theta, \sigma_{ri}^2)$. The posterior is obtained by multiplying the BFF with the prior density.

Figure 7 shows the BFFs related to a logistic regression analysis of the data including all exposures as main effects and an intercept term. Each BFF relates to the exponentiated regression coefficient, which can be interpreted as the multiplicative change of odds of neonatal death when increasing the variable by one unit while keeping all other variables fixed. An improper flat prior was assigned to the intercept under both the null and the alternative, while independent $p(\log \mathrm{OR}_i \mid H_1) = \mathrm{N}(\log \mathrm{OR}_i \mid 0, 1/2)$ priors were assigned to each coefficient under the alternative. This prior repre-

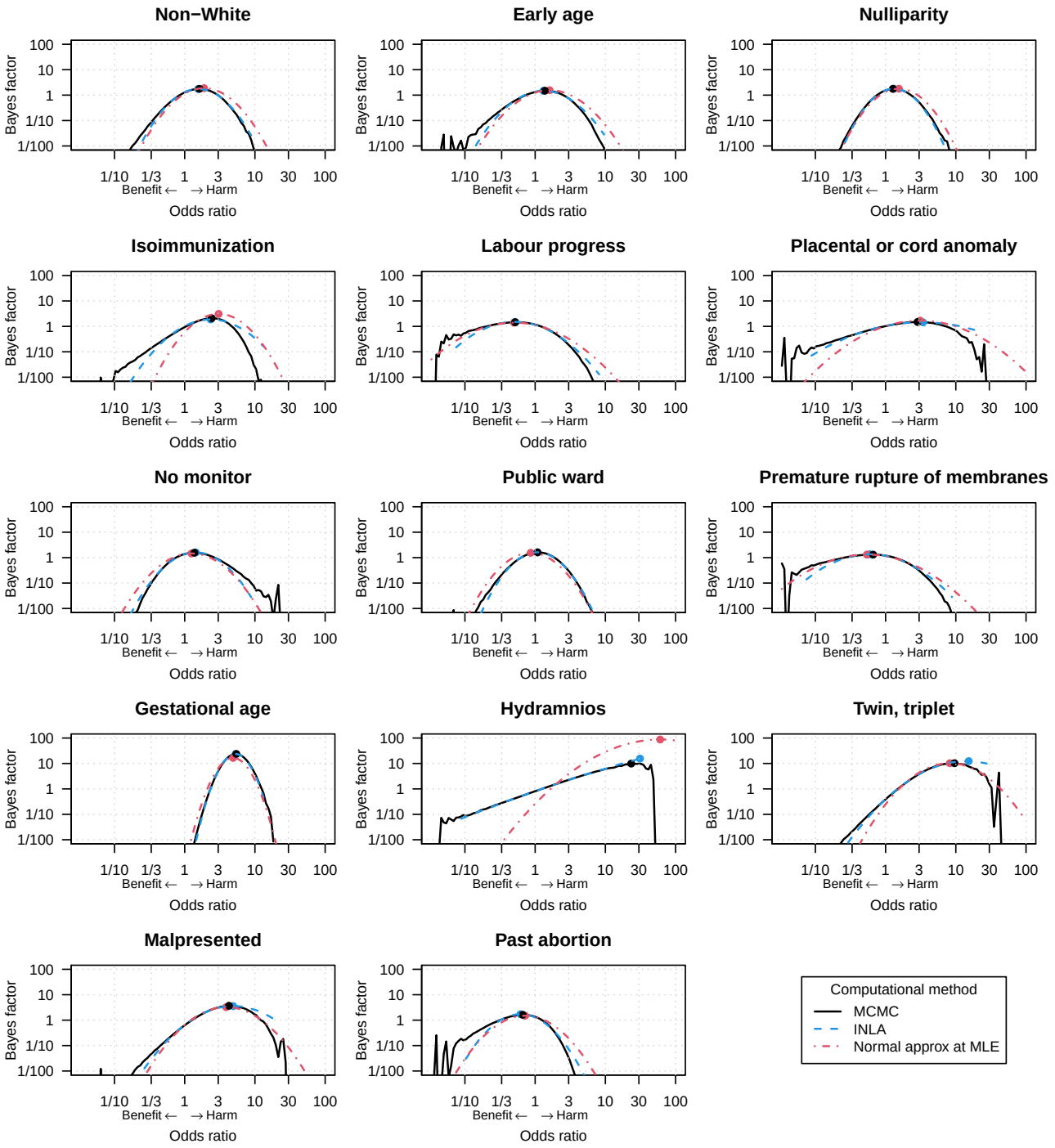

Figure 7: Multiple logistic regression BFF analysis of 2992 births with 17 neonatal deaths from Neutra et al. (1978) as reanalyzed by Sullivan and Greenland (2012). Each plot shows the BFF related to the exponential of a regression coefficient which can be interpreted as odds ratio. Independent, weakly informative priors $N(0, 1/2)$ are assigned to the coefficients under the alternative H_1, and when the coefficient are considered as nuisance parameters under the null H_0. Variables are indicators except early age (0, 20 ,1 15 19, 2 under 15), gestational age ($0 =$ no, $1 = 36 - 38$ weeks, $2 = 33 - 36$ weeks; under 33 weeks excluded), isoimmunization ($0 =$ no, $1 =$ Rh, $2 =$ ABO), labour progress ($0 =$ no, $0.33 =$ prolonged, $0.67 =$ protracted, $1 =$ arrested) and past abortion ($0 =$ none, $1 = 1, 2 = 2+$).

sents an alternative postulating that the median odds ratio is 1 and that 95% of odds ratios are in between 1/4 and 4 (Greenland, 2006). For the analysis of each coefficient, all other coefficients were considered as nuisance parameters with the same priors assigned to them under the null as under the alternative.

BFFs were computed in three ways: i) By first computing the marginal posterior distribution for each coefficient from 1'000'000 Markov chain Monte Carlo (MCMC) samples (solid black line) as implemented in Stan (Carpenter et al., 2017) and then computing the BFF via the Savage-Dickey density ratio. ii) Computing the marginal posterior with intergrated nested Laplace approximation (dashed blue line) as implemented in INLA (Rue et al., 2009) and then computing the BFF via the Savage-Dickey density. iii) The logistic model was first estimated with maximum likelihood, and each estimated coefficient and its standard error were then used for a univariate normal analysis as explained in Section 2.3 ignoring the nuisance parameters. The MCMC analysis took the longest among the three (several minutes), folled by the INLA analysis (several seconds), and the univariate analysis (almost instantly).

Due to the sparse nature of the data, most of the BFFs are relatively uninformative about whether or not the variables exhibit harmful or beneficial associations with neonatal death. For example, the BFF for the coefficient 'Non-White' (top left panel) is peaked at $\widehat{\text{OR}}_{\text{ME}} = 1.6$ indicating a slightly harmful association between non-white ethnicity and neonatal death, yet this parameter value receives only anecdotal support over the alternative ($k_{\text{ME}} = 1.76$) and the corresponding $k = 1$ support interval spans the range from slightly beneficial up to harmful association.

For most variables, extremely harmful or extremely beneficial associations are clearly ruled. However, the variables 'Gestational age', 'Hydramnios', and 'Twin, triplet' are notable exceptions. In each case, the BFFs suggest small up to very harmful associations with neonatal death. The most extreme among them is 'Hydramnios' for which an MEE of $\widehat{\text{OR}}_{\text{ME}} = 23.3$ is obtained. Due to the marginalization over the nuisance parameters, this estimate is somewhat smaller than the maximum likelihood estimate $\widehat{\text{OR}}_{\text{ML}} = 60.3$, yet still unrealistically large. This extreme inflation reflects the fact that only one death with hydramnios during pregnancy was observed.

5 Discussion

We showed how Bayes factors can be used for parameter estimation, extending their traditional use cases of hypothesis testing and model comparison. We also linked these ideas to the overarching concept of Bayes factor functions (BFFs), which are Bayes factor analogues of *P*-value functions, and are likewise particularly useful for reporting of analysis results. This provides data analysts with a unified framework for statistical inference that is distinct from conventional frequentist and Bayesian approaches: BFF inference uses the Bayesian evidence calculus, but without synthesizing data and prior. At the same time, BFF inference is closely related to likelihood-based inference, but also includes a natural way to deal with nuisance parameters.

Like the likelihoodist and Neyman-Pearson paradigms of statistical inference, BFF inference requires the formulation of alternative hypotheses. For this reason, BFFs are particularly valuable in contexts where there are prior data available or strong theories to formulate alternative hypotheses. For instance, BFFs (under the name of '*K* ratio') have been applied by the large-scale NANOGrav

collaboration to quantify the evidence for new physics theories against the established standard model (Afzal et al., 2023). In cases where there are no clear alternative hypotheses, data analysts may use BFFs based on 'weakly informative' (Gelman, 2009) or 'default' prior distribution (e.g., unit-information priors, see Kass and Wasserman, 1995) but should acknowledge this limitation and report sensitivity analyses (e.g., BFFs for different prior distributions). Another possibility is to base BFF inferences on Bayes factor bounds (Berger and Sellke, 1987; Sellke et al., 2001; Held and Ott, 2018), which give a bound on the maximum evidence against parameter values, but at the cost of losing the ability to quantify evidence *in favour* of parameter values (Pawel et al., 2023b).

Where under their control, data analysts should design experiments and studies so that conclusive inferences can be drawn from the data collected, and this is also applies to BFF inferences. Future research needs to investigate how experiments need to be designed to enable conclusive inference with BFFs. Finally, the computation of BFFs can be computationally demanding. For example, when a BFF is computed via the Savage-Dickey density from a posterior distribution computed by MCMC, the BFF at the tails of the posterior might be imprecise even with millions of samples. Future work may focus on developing more efficient techniques for computing BFFs in such settings.

Bayesian, likelihoodist, or predictive reasoning may all motivate the Bayes factor as a natural tool for quantifying the relative evidence or support of competing hypotheses. Nevertheless, neither the Bayes factor nor any other measure of statistical evidence is infallible or suitable for all purposes. Any type of statistical inference can lead to distorted scientific inferences when used in a bright-line fashion without consideration of contextual factors (Goodman, 2016; Greenland, 2023). We believe that BFFs are useful in this regard because they shift the focus from finding evidence against a single null hypothesis to making gradual and quantitative inferences.

Acknowledgments

We thank František Bartoš, Andrew Fowlie, Małgorzata Roos, Leonhard Held, and Eric-Jan Wagenmakers for valuable comments on drafts of the manuscript. The acknowledgment of these individuals does not imply their endorsement of the paper.

Conflict of interest

We declare no conflict of interest.

Software and data

The code and data to reproduce our analyses is openly available at `https://github.com/SamCH93/BFF`. A snapshot of the repository at the time of writing is available at `https://zenodo.org/doi/10.5281/zenodo.XXXXXX`. We used the statistical programming language R version 4.3.2 (2023-10-31) for analyses (R Core Team, 2023).

References

Afzal, A., Agazie, G., Anumarlapudi, A., Archibald, A. M., Arzoumanian, Z., Baker, P. T., Bécsy, B., Blanco-Pillado, J. J., Blecha, L., Boddy, K. K., Brazier, A., Brook, P. R., Burke-Spolaor, S., Burnette, R., Case, R., Charisi, M., Chatterjee, S., Chatziioannou, K., Cheeseboro, B. D., Chen, S., Cohen, T., Cordes, J. M., Cornish, N. J., Crawford, F., Cromartie, H. T., Crowter, K., Cutler, C. J., DeCesar, M. E., DeGan, D., Demorest, P. B., Deng, H., Dolch, T., Drachler, B., von Eckardstein, R., Ferrara, E. C., Fiore, W., Fonseca, E., Freedman, G. E., Garver-Daniels, N., Gentile, P. A., Gersbach, K. A., Glaser, J., Good, D. C., Guertin, L., Gültekin, K., Hazboun, J. S., Hourihane, S., Islo, K., Jennings, R. J., Johnson, A. D., Jones, M. L., Kaiser, A. R., Kaplan, D. L., Kelley, L. Z., Kerr, M., Key, J. S., Laal, N., Lam, M. T., Lamb, W. G., W. Lazio, T. J., Lee, V. S. H., Lewandowska, N., Lino dos Santos, R. R., Littenberg, T. B., Liu, T., Lorimer, D. R., Luo, J., Lynch, R. S., Ma, C.-P., Madison, D. R., McEwen, A., McKee, J. W., McLaughlin, M. A., McMann, N., Meyers, B. W., Meyers, P. M., Mingarelli, C. M. F., Mitridate, A., Nay, J., Natarajan, P., Ng, C., Nice, D. J., Ocker, S. K., Olum, K. D., Pennucci, T. T., Perera, B. B. P., Petrov, P., Pol, N. S., Radovan, H. A., Ransom, S. M., Ray, P. S., Romano, J. D., Sardesai, S. C., Schmiedekamp, A., Schmiedekamp, C., Schmitz, K., Schröder, T., Schult, L., Shapiro-Albert, B. J., Siemens, X., Simon, J., Siwek, M. S., Stairs, I. H., Stinebring, D. R., Stovall, K., Stratmann, P., Sun, J. P., Susobhanan, A., Swiggum, J. K., Taylor, J., Taylor, S. R., Trickle, T., Turner, J. E., Unal, C., Vallisneri, M., Verma, S., Vigeland, S. J., Wahl, H. M., Wang, Q., Witt, C. A., Wright, D., Young, O., and Zurek, K. M. (2023). The NANOGrav 15 yr data set: Search for signals from new physics. *The Astrophysical Journal Letters*, 951(1):L11. doi:10.3847/2041-8213/acdc91.

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(2):115–139. doi:10.1111/j.2517-6161.1949.tb00028.x.

Bartoš, F., Sarafoglou, A., Godmann, H. R., Sahrani, A., Leunk, D. K., Gui, P. Y., Voss, D., Ullah, K., Zoubek, M. J., Nippold, F., Aust, F., Vieira, F. F., Islam, C.-G., Zoubek, A. J., Shabani, S., Petter, J., Roos, I. B., Finnemann, A., Lob, A. B., Hoffstadt, M. F., Nak, J., de Ron, J., Derks, K., Huth, K., Terpstra, S., Bastelica, T., Matetovici, M., Ott, V. L., Zetea, A. S., Karnbach, K., Donzallaz, M. C., John, A., Moore, R. M., Assion, F., van Bork, R., Leidinger, T. E., Zhao, X., Motaghi, A. K., Pan, T., Armstrong, H., Peng, T., Bialas, M., Pang, J. Y. C., Fu, B., Yang, S., Lin, X., Sleiffer, D., Bognar, M., Aczel, B., and Wagenmakers, E.-J. (2023). Fair coins tend to land on the same side they started: Evidence from 350, 757 flips. doi:10.48550/ARXIV.2310.04153. arXiv preprint.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, 72(358):355–366. doi:10.1080/01621459.1977.10481002.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577. doi:10.1214/12-aos1013.

Bender, R., Berg, G., and Zeeb, H. (2005). Tutorial: Using confidence curves in medical research. *Biometrical Journal*, 47(2):237–247. doi:10.1002/bimj.200410104.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3):317–335. doi:10.1214/ss/1177013238.

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1). doi:10.1214/ss/1009211804.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, 82(397):112. doi:10.2307/2289131.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons, Hoboken. doi:10.1002/9780470316870.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association*, 56(294):246–249. doi:10.1080/01621459.1961.10482107.

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, 21(17):2563–2599. doi:10.1002/sim.1216.

Campbell, H. and Gustafson, P. (2022). Bayes factors and posterior estimation: Two sides of the very same coin. *The American Statistician*, 77(3):248–258. doi:10.1080/00031305.2022.2139293.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1). doi:10.18637/jss.v076.i01.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13(2). doi:10.1214/18-ba1103.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372. doi:10.1214/aoms/1177706618.

Dawid, P. A. (2011). Posterior model probabilities. In Bandyopadhyay, P. S. and Forster, M. R., editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 607–630. North-Holland, Amsterdam.

Diaconis, P., Holmes, S., and Montgomery, R. (2007). Dynamical bias in the coin toss. *SIAM Review*, 49(2):211–235. doi:10.1137/s0036144504446436.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42(1):204–223. doi:10.1214/aoms/1177693507.

Edwards, A. W. F. (1971). *Likelihood*. Cambridge University Press, London.

Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics - Theory and Methods*, 26(5):1125–1143. doi:10.1080/03610929708831972.

Evans, M. (2015). *Measuring statistical evidence using relative belief*. CRC Press, Boca Raton.

Fowlie, A. (2024). The Bayes factor surface for searches for new physics. doi:10.48550/ARXIV.2401.11710. arXiv preprint.

Franck, C. T. and Gramacy, R. B. (2019). Assessing Bayes factor surfaces using interactive visualization and computer surrogate modeling. *The American Statistician*, 74(4):359–369. doi:10.1080/00031305.2019.1671219.

Fraser, D. A. S. (2019). The *p*-value function and statistical inference. *The American Statistician*, 73(sup1):135–147. doi:10.1080/00031305.2018.1556735.

Gelman, A. (2009). Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science*, 24(2):176–178. doi:10.1214/09-sts284d.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. doi:10.1198/016214506000001437.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Griffin, London.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114. doi:10.1111/j.2517-6161.1952.tb00104.x.

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813. doi:10.1080/01621459.1958.10501480.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130(12):1005. doi:10.7326/0003-4819-130-12-199906150-00019.

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science*, 352(6290):1180–1181. doi:10.1126/science.aaf5406.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International Journal of Epidemiology*, 35(3):765–775. doi:10.1093/ije/dyi312.

Greenland, S. (2007). Bayesian perspectives for epidemiological research. ii. regression analysis. *International Journal of Epidemiology*, 36(1):195–202. doi:10.1093/ije/dyl289.

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology*, 186(6):639–645. doi:10.1093/aje/kwx259.

Greenland, S. (2023). Divergence versus decision *P*-values: A distinction worth making in theory and keeping in practice: Or, how divergence *P*-values measure evidence evfen when decision *P*-values do not. *Scandinavian Journal of Statistics*, 50(1):54–88. doi:10.1111/sjos.12625.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350. doi:10.1007/s10654-016-0149-3.

Grünwald, P., de Heide, R., and Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. doi:10.48550/arXiv.1906.07801.

Harms, C. (2019). A Bayes factor for replications of ANOVA results. *The American Statistician*, 73(4):327–339. doi:10.1080/00031305.2018.1518787.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385. doi:10.1093/biomet/61.2.383.

Held, L., Matthews, R., Ott, M., and Pawel, S. (2022). Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*, 13(3):295–314. doi:10.1002/jrsm.1538.

Held, L. and Ott, M. (2018). On *p*-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5(1):393–419. doi:10.1146/annurev-statistics-031017-100307.

Hendriksen, A., de Heide, R., and Grünwald, P. (2021). Optional stopping with Bayes factors: A categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, 16(3). doi:10.1214/20-ba1234.

Infanger, D. and Schmidt-Trucksäss, A. (2019). *P* value functions: An underused method to present research results and to promote quantitative reasoning. *Statistics in Medicine*, 38(21):4189–4197. doi:10.1002/sim.8293.

Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press, Oxford, first edition.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford, third edition.

Johnson, V. E., Pramanik, S., and Shudde, R. (2023). Bayes factor functions for reporting outcomes of hypothesis tests. *Proceedings of the National Academy of Sciences*, 120(8). doi:10.1073/pnas.2217331120.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2):175–194. doi:10.1111/j.2517-6161.1970.tb00830.x.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. doi:10.1080/01621459.1995.10476572.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934. doi:10.1080/01621459.1995.10476592.

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *The Annals of Mathematical Statistics*, 34(3):1109–1110. doi:10.1214/aoms/1177704038.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974. doi:10.2307/2529876.

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*, 51(6):2498–2508. doi:10.3758/s13428-018-1092-x.

Marschner, I. C. (2024). Confidence distributions for treatment effects in clinical trials: Posteriors without priors. *Statistics in Medicine*. doi:10.1002/sim.10000.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. National Academies Press. doi:10.17226/25303.

Neutra, R. R., Fienberg, S. E., Greenland, S., and Friedman, E. A. (1978). Effect of fetal monitoring on neonatal death rates. *New England Journal of Medicine*, 299(7):324–326. doi:10.1056/nejm197808172990702.

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 73(sup1):69–81. doi:10.1080/00031305.2018.1518265.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistic 2B*. Wiley & Sons, Chichester, second edition.

Pawel, S., Aust, F., Held, L., and Wagenmakers, E.-J. (2023a). Power priors for replication studies. *TEST*. doi:10.1007/s11749-023-00888-5.

Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):879–911. doi:10.1111/rssb.12491.

Pawel, S., Ly, A., and Wagenmakers, E.-J. (2023b). Evidential calibration of confidence intervals. *The American Statistician*, pages 1–11. doi:10.1080/00031305.2023.2216239.

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., and Schooler, J. W. (2023). High replicability of newly discovered social-behavioural findings is achievable. *Nature Human Behaviour*. doi:10.1038/s41562-023-01749-9.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rafi, Z. and Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, 20:244. doi:10.1186/s12874-020-01105-9.

RECOVERY collaborative group (2022). Baricitinib in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial and updated meta-analysis. *The Lancet*, 400(10349):359–368. doi:10.1016/s0140-6736(22)01109-6.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409. doi:10.1214/aoms/1177696786.

Rosenthal, R. and Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, 5(6):329–334. doi:10.1111/j.1467-9280.1994.tb00281.x.

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., and Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, 12(4):448–474. doi:10.1002/jrsm.1475.

Royall, R. (1997). *Statistical Evidence: A likelihood paradigm*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392. doi:10.1111/j.1467-9868.2008.00700.x.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability*. Cambridge University Press, Cambridge.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71. doi:10.1198/000313001300339950.

Severini, T. A. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika*, 94(3):529–542. doi:10.1093/biomet/asm040.

Shalloway, D. (2014). The evidentiary credible region. *Bayesian Analysis*, 9(4). doi:10.1214/14-ba883.

Stone, M. (1997). Discussion of papers by Dempster and Aitkin. *Statistics and Computing*, 7(4):263–264. doi:10.1023/a:1018502622516.

Sullivan, S. G. and Greenland, S. (2012). Bayesian regression in sas software. *International Journal of Epidemiology*, 42(1):308–317. doi:10.1093/ije/dys213.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618. doi:10.1080/01621459.1995.10476554.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457–1475. doi:10.1037/a0036731.

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F., and Etz, A. (2020). The support interval. *Erkenntnis*. doi:10.1007/s10670-019-00209-z.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60(3):158–189. doi:10.1016/j.cogpsych.2009.12.001.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39. doi:10.1111/insr.12000.

Computational details

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2024-02-21 15:51:01.894987 Europe/Zurich

sessionInfo()

## R version 4.3.2 (2023-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.10.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Zurich
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] INLA_24.02.09            sp_2.1-3                 Matrix_1.6-3
## [4] metabf_0.1               ReplicationSuccess_1.3.1 metadat_1.3-0
## [7] knitr_1.44
##
## loaded via a namespace (and not attached):
##  [1] dplyr_1.1.4        compiler_4.3.2     highr_0.10         tidyselect_1.2.0
##  [5] Rcpp_1.0.12        MatrixModels_0.5-3 parallel_4.3.2     splines_4.3.2
##  [9] lattice_0.22-5     R6_2.5.1           generics_0.1.3     classInt_0.4-10
## [13] sf_1.0-14          tibble_3.2.1       fmesher_0.1.5      units_0.8-4
## [17] DBI_1.1.3          pillar_1.9.0       rlang_1.1.1        utf8_1.2.3
## [21] mathjaxr_1.6-0     xfun_0.40          cli_3.6.1          withr_2.5.0
```

```
## [25] magrittr_2.0.3     class_7.3-22       digest_0.6.33      grid_4.3.2
## [29] lifecycle_1.0.3    vctrs_0.6.5        KernSmooth_2.23-22 proxy_0.4-27
## [33] evaluate_0.21      glue_1.6.2         fansi_1.0.4        e1071_1.7-13
## [37] tools_4.3.2        pkgconfig_2.0.3
```