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

Version August 29, 2025

Abstract

The Bayes factor, the data-based updating factor of the prior to posterior odds of two hypotheses, is a natural measure of statistical evidence for one hypothesis over the other. 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 'support curve' is 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 as Bayes factors (for any tested parameter value), support intervals (at any level), and point estimates can be easily read off from a plot of the support curve. This approach shares similarities but is also distinct from conventional Bayesian and frequentist approaches: It uses the Bayesian evidence calculus, but without synthesizing data and prior, and it defines statistical evidence in terms of (integrated) likelihood ratios, but also includes a natural way for dealing with nuisance parameters. Applications to meta-analysis, replication studies, and logistic regression illustrate how our framework is of practical value for making 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 region 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 distinct, 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 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 (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(the treatment effect is absent)' and 'Pr(the treatment effect is present)'.

One approach to address this issue 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

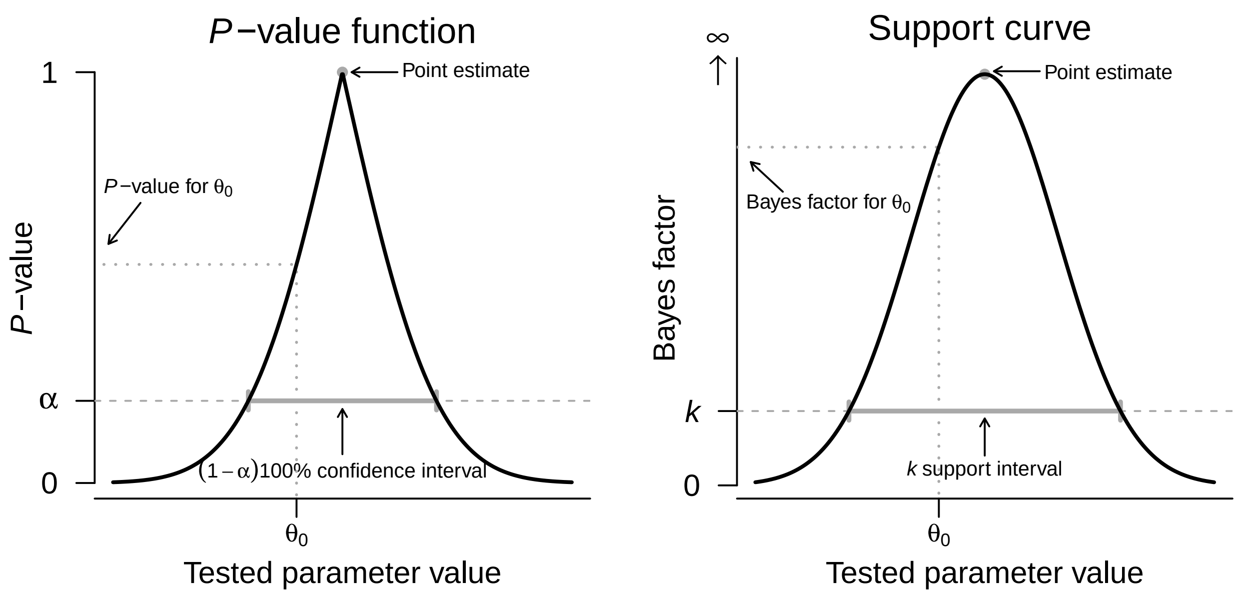

Figure 1: Examples of P-value functions and support curves. P-value are two-sided. Bayes factors are oriented in favour of the tested parameter value over a specified alternative hypothesis (i.e., a higher Bayes factor indicates higher support for the parameter value over the alternative).

such, Bayes factors allow data analysts to evaluate the relative evidence for two hypotheses without depending on the prior probabilities of the hypotheses; 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., 2022; Kelter, 2022), or they can assign prior probabilities to the tested hypotheses and report a posterior averaged over both hypotheses (known as Bayesian model averaging, see e.g., Hoeting et al., 1999; Rouder et al., 2018; Campbell and Gustafson, 2022), but this requires specification of prior probabilities which is highly controversial and usually 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 resolve this dilemma and provide a unified framework for estimation and hypothesis testing based on Bayes factors. 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 *support curve* to derive point estimates, interval estimates, and Bayes factors (as shown in the right plot in Figure 1), establishing a duality between hypothesis testing and parameter estimation. This provides data analysts with a unified framework for statistical inference centred around the Bayes factor.

Interval estimates based on the Bayes factor have been proposed under the name of the Bayesian support interval (Wagenmakers et al., 2022; Pawel et al., 2024). Similarly, looking at the Bayes factor as a function of parameter values has been proposed in the physics community under the names of 'Bayes factor surface' (Fowlie, 2024) and 'K ratio' (Afzal et al., 2023), and by Rouder and Morey (2018) in the context of teaching Bayesian statistics. A method called 'Bayes factor function' has been proposed by Johnson et al. (2023). In this approach, Bayes factors are viewed as a function of a hyperparameter of the prior *under the alternative hypothesis* but for a fixed null hypothesis. The Bayes factor function can in some cases be made equivalent to the support curve if the roles of the null and alternative hypothesis are reversed and point mass priors are assigned to the parameter under the alternative, yet this seems rather unnatural from the perspective of both methods. Finally, a similar attempt of reconciling Bayesian interval estimation and hypothesis testing with a focus on the 'Bayesian evidence value' has been investigated by Kelter (2022).

While many elements of support curve inference have been considered in earlier articles, this paper provides a synthesis of these ideas from the perspective of the support curve as a counterpart to the P-value function. A novel contribution is the concept of point estimation based on Bayes factors. We call the resulting estimate the *maximum evidence estimate* (MEE). The MEE is the parameter value that maximizes the support curve, and as such, is the parameter value that receives the most evidential support from the data over a specified alternative hypothesis. Finally, this paper shows how support curves can be applied in several areas where they have never been used before (meta-analysis, replication studies, and logistic regression).

It is important to note that our intent is not to fuel the 'statistics wars' and entertain the position that Bayesian inference or Bayes factors are the best or only valid form of statistical inference. The goal is simply to explore the idea of a Bayes factor analogue of the P-value function, find out how such an approach can be useful, and how it connects to other approaches. Bayes factors, just as P-values and posterior probabilitites, can be delicate tools and users should be wary of their limitations, misuses, and misinterpretations (Tendeiro and Kiers, 2019; Wong et al., 2022; Tendeiro et al., 2024).

This paper is structured as follows. In the following (Section 2), we introduce the theoretical foundations of Bayes factors, support sets, and maximum evidence estimates. We then explore their connection to other statistical frameworks (Section 3). Real data examples from meta-analysis, replication studies, and logistic regression then illustrate properties of the Bayes factor framework (Section 4). We conclude with a discussion of limitations, advantages, and directions for future research (Section 5).

2 Support curve inference

Suppose we observe data y with an assumed distribution with probability density/mass function $p(y \mid \theta, \psi)$ that depends on parameters $\theta \in \Theta$ and $\psi \in \Psi$, with θ being the focus parameters (the parameters of substantial interest) 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 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

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

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

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

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. Note that Bayes factors do not require the null and alternative models to be nested but are applicable to general model comparisons. However, due to their more intuitive interpretation and computational advantages, we will focus on nested models and discuss extensions to non-nested models in Section 5.

As (1a) shows, the Bayes factor represents the data-based core of the Bayesian belief calculus since it transforms the prior odds of H_0 versus H_1 to the corresponding posterior odds (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). Since the Bayes factor remains the same for any pair of prior hypothesis probabilities $p(H_0)$ and $p(H_1)$, it remains useful to different analysts with different prior hypothesis probabilities, or even if one rejects the idea of assigning probabilities to H_0 and H_1. However, this does not mean that the Bayes factor does not depend on any prior distributions; in fact, it can be quite sensitive to the priors assigned to the parameters ψ (and θ) under H_0 and H_1, respectively (Kass and Raftery, 1995). We will return to the choice of parameter priors in Section 2.4. 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 integrated over the prior of ψ under H_0) by the likelihood of y integrated over the prior of θ (and possibly ψ) under H_1. The priors for θ and ψ may also be point mass 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 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 *support curve* (SC) 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 SC, 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; Bind and Rubin, 2024).

For one- or two-dimensional focus parameters θ, the SC 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 (for example, more than two regression coefficients), this becomes more difficult. In this case, some focus parameters can be re-considered as nuisance parameters to reduce the SC visualization again to a curve or surface. This is similar to ordinary Bayesian inference where a marginal posterior is obtained from marginalizing a joint posterior.

2.1 Support sets

Apart from visual inspections, we may also want to summarize the SC numerically. The SC can be used to obtain *support sets* (Wagenmakers et al., 2022) 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

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

that is, the parameter values for which the Bayes factor indicates evidence of at least level k over the specified alternative. A k support set thus contains all the parameter values that, when considered as individual null hypotheses compared to an alternative, receive a Bayes factor of at least k. Each included parameter value is therefore associated with an updating factor of at least k or, equally, relative predictive performance of at least k. Support sets are thus 'calibrated' with respect to prior-posterior updating, as opposed to the calibration of confidence sets with respect to long-run coverage, or that of posterior credible intervals with respect to posterior belief.

In practice, a k support set (typically an interval) is obtained from 'cutting' the SC 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 convention 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 SC is given by

$$\hat{\theta}_{\mathrm{ME}} = \operatorname*{arg\,max}_{\theta_0 \in \Theta} \mathrm{BF}_{01}(y;\theta_0),$$

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*

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

that is, the SC 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 SC is hence given by 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 the behaviour of the MEE with increasing sample size, we may look at an approximation of the Bayes factor. Suppose that the data $y_{1:n} = \{y_1, y_2, \dots, y_n\}$ 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 $n\widehat{V}_0$ and $n\widehat{V}_1$ 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 SC (O'Hagan and Forster, 2004, equation 7.27) gives then

$$\log \text{BF}_{01}(y_{1:n}; \theta_0) \approx \log \frac{p(y_{1:n} \mid \theta_0, \hat{\psi}_0)}{p(y_{1:n} \mid \hat{\theta}_1, \hat{\psi}_1)} + \frac{\dim(\theta)}{2} \log \frac{n}{2\pi} + \log \frac{p(\hat{\psi}_0 \mid H_0)}{p(\hat{\theta}_1, \hat{\psi}_1 \mid H_1)} + \frac{1}{2} \log \frac{|\widehat{V}_0|}{|\widehat{V}_1|}. \quad (2)$$

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 SC 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 SC. 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 approximately to maximizing the profile likelihood that is adjusted based on the accuracy of the prior of the parameters and the modal dispersion.

2.3 Example: Normal mean

Suppose we observe a mean $\bar{y}$ from n observations, assumed to be sampled (at least approximately) from a normal distribution $\bar{Y} \mid \theta \sim \text{N}(\theta, \kappa^2/n)$. Assume that the variance κ^2 is known and we want to conduct inferences regarding θ. For example, suppose we observed a mean test score of $\bar{y} = 110$

in a sample of $n = 30$ students. Also assume that in the general population the average score is 100 and the standard deviation is $\kappa = 15$. 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.

Table 1: Support curve BF_{01}, maximum evidence estimate $\hat{\theta}_{\mathrm{ME}}$, evidence value k_{ME}, and k support interval (SI) for a normal mean based on an observed mean $\bar{y}$ from $\bar{Y} \mid \theta \sim \mathrm{N}(\theta, \kappa^2/n)$ with known variance κ^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		
	$\theta \sim \mathrm{N}(m, v)$	$\theta \sim \mathrm{N}(\theta_0, v)$	$\theta = \theta_0 + d$
BF_{01}	$\exp\left[-\frac{1}{2}\left\{\frac{(\bar{y}-\theta_0)^2}{\kappa^2/n} - \frac{(\bar{y}-m)^2}{v+\kappa^2/n}\right\}\right]\sqrt{1+\frac{v\,n}{\kappa^2}}$	$\exp\left[-\frac{1}{2}\left\{\frac{n(\bar{y}-\theta_0)^2}{\kappa^2\{1+\kappa^2/(v\,n)\}}\right\}\right]\sqrt{1+\frac{v\,n}{\kappa^2}}$	$\exp\left\{\frac{2d(\theta_0-\bar{y})+d^2}{2\kappa^2/n}\right\}$
$\hat{\theta}_{\mathrm{ME}}$	$\bar{y}$	$\bar{y}$	non-existent
k_{ME}	$\exp\left\{\frac{(\bar{y}-m)^2}{2(v+\kappa^2/n)}\right\}\sqrt{1+\frac{v\,n}{\kappa^2}}$	$\sqrt{1+\frac{v\,n}{\kappa^2}}$	non-existent
k SI	$\bar{y} \pm \frac{\kappa}{\sqrt{n}}\sqrt{\log(1+\frac{v\,n}{\kappa^2}) + \frac{(\bar{y}-m)^2}{v+\kappa^2/n} - \log k^2}$	$\bar{y} \pm \frac{\kappa}{\sqrt{n}}\sqrt{\{\log(1+\frac{v\,n}{\kappa^2}) - \log k^2\}\,(1+\frac{\kappa^2}{v\,n})}$	$\left[\bar{y} + \frac{\kappa^2 \log k}{n\,d} - \frac{d}{2}, \infty\right)$

Perhaps the simplest choice is a prior that does not depend on the parameter value θ_0 of the null hypothesis, such as a normal prior with mean m and variance v (left column in Table 1). The hyper-parameters m and v may be specified based on external data or based on an alternative hypothesis of interest. For example, suppose that 50 students from the previous year had a mean test score of 120, which we now use to set the prior mean and variance to $m = 120$ and $v = 15^2/50$, see Figure 2 for the resulting SC (orange). In this case, the MEE is given by $\hat{\theta}_{\mathrm{ME}} = \bar{y} = 110$ 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 a wide range from 101.9 to 118.1, indicating that average up to above average test scores 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 observed 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, and thus fewer point null hypotheses can outpredict it. Figure 2 illustrates this phenomenon with another prior distributions (one with mean at the observed test score $\bar{y} = 110$, the blue SC). The SC has its mode still at the observed mean test score but shows a far narrower

$k = 1$ support interval from 108.1 to 111.9 than the orange SC with the mean $m = 120$ based on the previous year students.

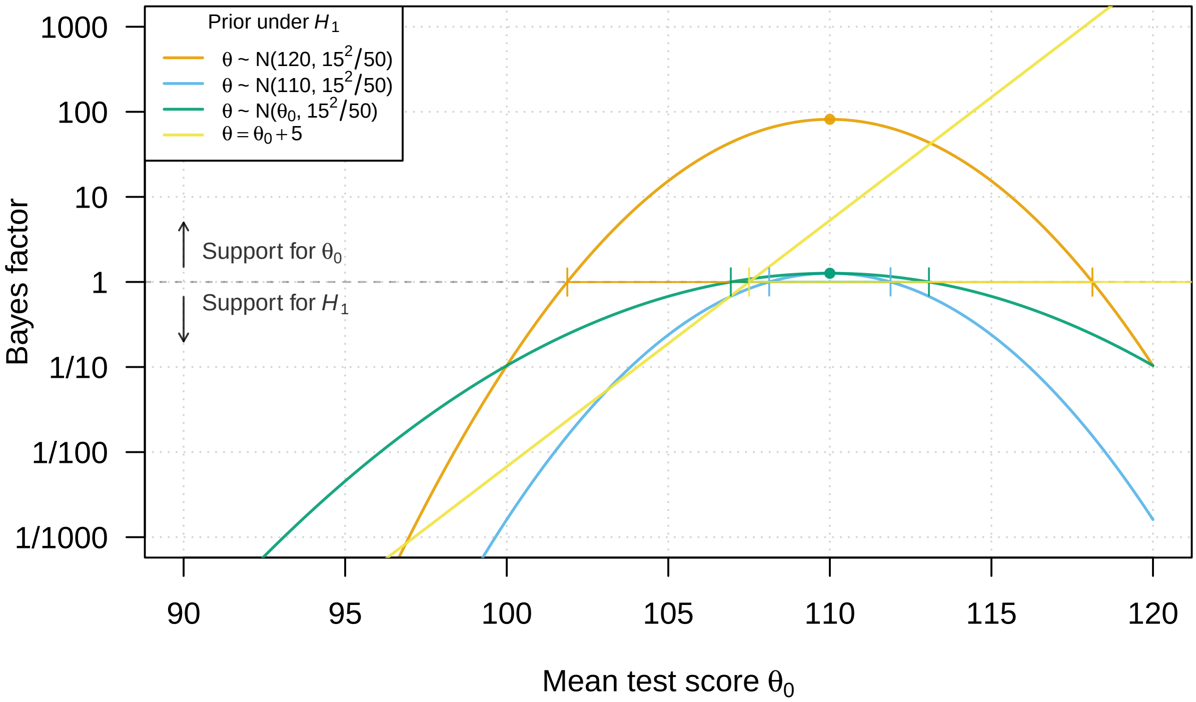

Figure 2: Support curve, MEE, and $k = 1$ support interval for a mean test score θ based on the observed mean $\bar{y} = 110$ from $n = 30$ observations for different prior distributions for θ under the alternative H_1. A normal likelihood $\bar{Y} \mid \theta \sim \mathrm{N}(\theta, 15^2/n)$ is assumed for the data.

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, the resulting SC varies both the null and the alternative, unlike the SC based on the 'global' normal prior with fixed mean m. As a result, the interpretation of the SC is different: For such a 'local' normal prior, the SC 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 $\hat{\theta}_{\mathrm{ME}} = y$ and support intervals are centered around it, but the associated, Bayes factor, evidence level and support interval are different. Figure 2 illustrates that when the mean m of a global normal prior is too different from the observed mean $\bar{y}$ (as in the case of the orange SC, where the prior was specified based on the students from the previous year), 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 SC), 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 the alternative hypothesis that the mean test score is greater than the tested parameter value. However, this leads to an ever-increasing SC, see Figure 2 for an illustration. As a result, the MEE and its evidence level do not exist, while 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 SC.

2.4 Choice of the prior

While Bayes factors do not depend on the prior probabilities of their contrasted hypothesis, they can be sensitive to the prior distributions specified for model parameters under each hypothesis. Consequently, the parameter priors also have a substantial impact on SC inferences, as the previous example showed. Moreover, unlike posterior distributions, where the influence of the prior diminishes as data accumulate, the sensitivity of Bayes factors to prior distributions persists even with large samples (Kass and Raftery, 1995).

This sensitivity is not inherently a weakness. It enables data analysts to accurately quantify the support of parameter values over informative alternative hypotheses when they are available (Vanpaemel, 2010), but poses a challenge in their absence (Liu and Aitkin, 2008; Williams et al., 2023). Even when external information exists, translating it into a prior distribution is often nontrivial. For instance, knowing that a parameter lies within a certain range still leaves considerable flexibility in the actual specification of the prior.

Several strategies have been proposed to address these challenges, each with advantages and limitations. First, so-called 'default' or 'objective' priors aim to yield Bayes factors with desirable properties, such as model selection consistency, without incorporating prior knowledge (Bayarri et al., 2012; Consonni et al., 2018). These are often provided as defaults in software for computing Bayes factors, such as the `BayesFactor` R package (Morey and Rouder, 2024) or `JASP` (Love et al., 2019). Second, prior elicitation techniques can be used to formally translate expert judgments into quantitative priors (O'Hagan, 2019; Stefan et al., 2022; Mikkola et al., 2024). Third, Bayes factor bounds (Berger and Sellke, 1987; Sellke et al., 2001; Held and Ott, 2018) and sensitivity analysis (Kass and Raftery, 1995;

Sinharay and Stern, 2002; Franck and Gramacy, 2019) explore how Bayes factors change across a class of priors to assess robustness of conclusions. Finally, reverse-Bayes analyses identify the priors that lead to the Bayes factor crossing a tipping point, allowing analysts to assess the plausibility of such a prior in light of external knowledge (Good, 1950; Held et al., 2022). As all of these techniques are applicable to Bayes factors, they are also applicable to SCs.

As in other Bayes factor applications, SCs are only unambiguously defined if priors for focus parameters are proper (i.e., integrate to one) under the alternative H_1. Assigning improper or very vague priors can lead SCs to diverge to infinity thereby proving overwhelming support for every tested parameter value θ_0, which is known as the 'Jeffreys-Lindley paradox' (Robert, 2014; Wagenmakers and Ly, 2023). 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* and *local* priors. Global priors do not depend on the parameter value θ_0 of the null hypothesis. In contrast, local priors are defined relative to θ_0. For example, a local prior may be centered around θ_0 or shifted away from it, as the priors underlying the green and yellow SCs in Figure 2. This results in a different alternative hypothesis for each tested parameter value, making the interpretation of the SC more intricate. 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 used in the same situations where default Bayes factors would be used.

Given these considerations, SCs, just as Bayes factors, are especially useful if informative priors based on prior data or external knowledege are available. If such priors are lacking and default priors are used, it is advisable to report sensitivity analyses for plausible ranges of priors or reverse-Bayes analysis, to assess the robustness of the conclusions. A convenient visual sensitivity analysis is, for example, to plot different SCs resulting from different prior specifications, as shown in Figure 2.

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 SC based on both batches is

$$\mathrm{BF}_{01}(y_{1:2};\theta_0) = \mathrm{BF}_{01}(y_1;\theta_0) \times \mathrm{BF}_{01}(y_2 \mid y_1;\theta_0)$$

where

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

is the *partial Bayes factor* obtained from using the posterior distributions under the null $p(\psi \mid y_1, H_0)$ and under the alternative $p(\theta, \psi \mid y_1, H_1)$ 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

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

that is, a SC based on all the available data can be obtained by multiplying the SC based on the previous batches by the partial Bayes factor based on the current batch. Thus, like ordinary Bayesian inference with posterior distributions, SC inference is sequentially coherent. In practice, it may, however, not always be computationally straightforward to reuse posterior distributions as prior distributions for subsequent analyses, particularly in high-dimensional settings.

2.6 Asymptotic behaviour of the support curve

It is of interest to understand the asymptotic behaviour of the SC, that is, how does the SC (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., Kass and Vaidyanathan, 1992; Gelfand and Dey, 1994; Dawid, 2011). Since the SC is nothing else than the Bayes factor evaluated for various null hypotheses, this consistency property carries over to the SC. That is, as more data are generated from the true model with parameter θ_*, the SC at $\theta_0 = \theta_*$ will go to infinity, while the SC at $\theta_0 \neq \theta_*$ will go to zero.

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

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

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

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

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

$$\Pr\{\mathrm{BF}_{01}(\bar{Y}; \theta_0) \leq \gamma \mid \theta_*\} = 1 - \Pr(\chi^2_{1,\lambda} \leq X) \tag{4}$$

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 SC 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 SC 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 SC 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 SC inference to other inference frameworks.

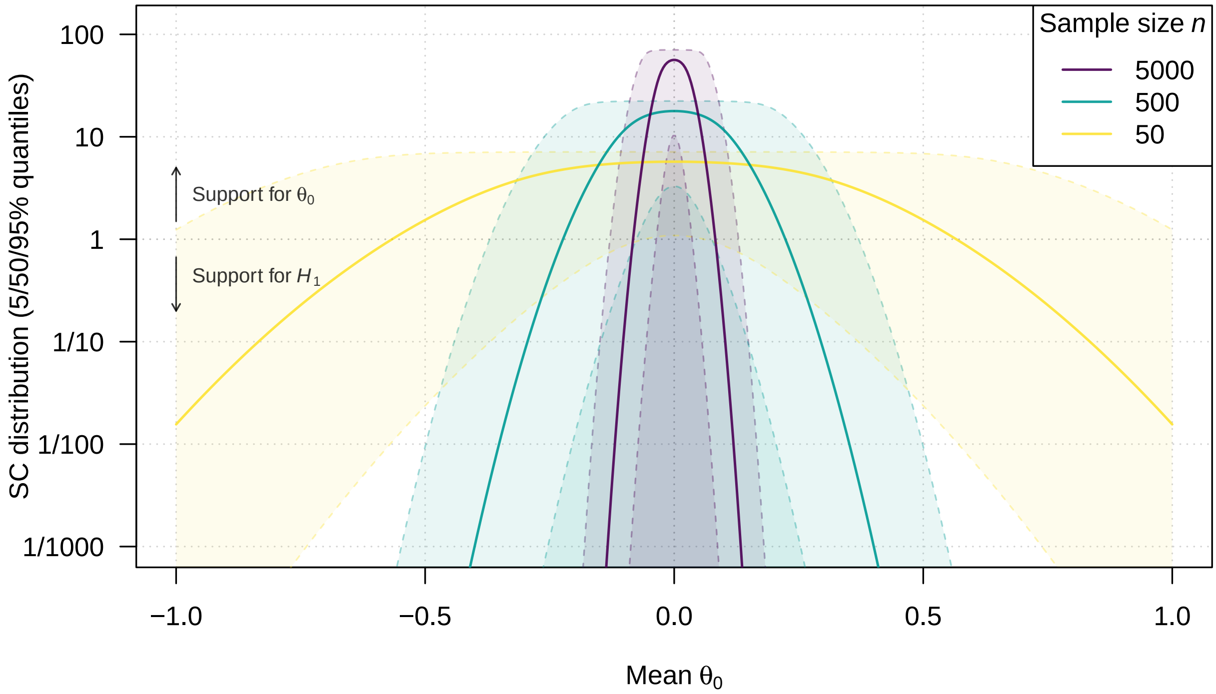

Figure 3: The distribution of the SC, summarized by 5% quantile (lower dashed line), 50% quantile (solid line), 95% quantile (upper dashed line), for different sample sizes n (colors). A data model $\bar{Y} \mid \theta \sim \mathrm{N}(\theta, \kappa^2/n)$ is assumed with true mean $\theta_* = 0$ and unit-variance $\kappa^2 = 4$. The SC is based on a local normal prior $\theta \mid H_1 \sim \mathrm{N}(\theta_0, v = 4)$ assigned to θ under the alternative. Quantiles were computed by numerically inverting the cumulative distribution function of the Bayes factor (4).

3.1 Maximum integrated likelihood

In typical situation where a division of the null's marginal likelihood $p(y \mid H_0)$ by the alternative's marginal likelihood $p(y \mid H_1)$ does not change the maximizer of the null's marginal likelihood $p(y \mid H_0)$, the MEE can be obtained by maximizing $p(y \mid H_0)$ without reference 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*

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

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 $Y_1, \dots, Y_n \mid$

$\theta, \sigma^2 \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$. 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 $y_1, \dots, y_n$ is then

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

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

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

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 and MIL estimates can also be motivated from a Bayes factor 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 (integrated) maximum likelihood estimate 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 SC 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, SC 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

$$L(\theta) = \frac{p(y \mid \theta)}{p(y \mid \hat{\theta}_{\mathrm{ML}})} \tag{5}$$

where $\hat{\theta}_{\text{ML}}$ is the maximum likelihood estimate. 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 maximum likelihood estimate 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 (5) 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 SC inference is given by the 'universal bound' (Kerridge, 1963; Robbins, 1970; Royall, 1997; Sanborn and Hills, 2013), which bounds the frequentist probability of obtaining misleading Bayesian evidence. That is, when data are generated under $H_0 : \theta = \theta_0$, the probability of obtaining misleading evidence against the null at level k (with a $\text{BF}_{01} \leq k < 1$) is at most k for any prior under the alternative

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

This can be shown by noting that under typical conditions the expectation of BF_{10} under H_0 is 1 and applying Markov's inequality (for technical details see e.g., Grünwald et al., 2024). 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 SCs into conservative P-values and confidence sets, e.g., a $k = 1/20$ support set obtained from a SC corresponds to a 95% conservative confidence set and $p = \min\{\text{BF}_{01}, 1\}$ corresponds to a conservative P-value. Remarkably, the bound holds without adjustment even in 'optional stopping' settings where 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 invalid.

3.4 Bayesian inference

The SC can, under certain conditions, be transformed into a Bayesian posterior distribution. Specifically, assuming that the priors for the nuisance parameters satisfy

$$p(\psi \mid H_0) = p(\psi \mid \theta = \theta_0, H_1), \tag{6}$$

the Bayes factor can be represented as the ratio of marginal posterior to prior density under H_1 evaluated at the tested parameter value (known as Savage-Dickey density ratio, see e.g., Dickey, 1971; Verdinelli and Wasserman, 1995; Wagenmakers et al., 2010). Hence, the posterior under H_1 can be obtained by multiplying the SC with the prior

$$p(\theta \mid y, H_1) = \underbrace{\frac{p(y \mid \theta, H_1)}{p(y \mid H_1)}}_{=\mathrm{BF}_{01}(y;\theta)} \times p(\theta \mid H_1). \tag{7}$$

It is, however, important to emphasize that SCs 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.

Since, under certain regularity conditions, the posterior is asymptotically normally distributed around the maximum likelihood estimate (Bernardo and Smith, 2000, chapter 5.3), we can conclude that whenever these conditions are satisfied and the SC has the Savage-Dickey density ratio representation (7), the SC is asymptotically given by the asymptotic posterior normal density divided by the prior density, both evaluated at θ_0. The posterior, and hence also the SC, will become more concentrated around the true parameter θ_* as more data are generated, giving another intuition about the consistency property of Bayes factors.

The Savage-Dickey density ratio (7) also provides a convenient way to compute SCs: 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 SC. The caveat is again that this only works for global priors under the alternative and provided that condition (6) holds for the priors assigned to the nuisance parameters.

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

$$\mathrm{RB}(\theta \mid H_1) = \underbrace{\frac{p(\theta \mid y, H_1)}{p(\theta \mid H_1)}}_{=\mathrm{BF}_{01}(y;\theta)}, \tag{8}$$

see e.g., Evans (2015). This quantity is the updating factor of the prior to the posterior density/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 is equivalent to the MEE whenever the SC and relative belief ratio coincide. Evans (1997) also defined a $\delta \times 100\%$ *relative surprise region*, which is the set of parameter values with $\delta \times 100\%$ 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 evidence, the ordering induced by the relative belief ratio merely provides a rule to chose among all credible sets (Wagenmakers et al., 2022). 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 k *plausible region* which contains parameter values with a relative belief ratio of at least k and as such coincides with the k support set whenever the Savage-Dickey density representation applies to the SC and when a global prior is chosen for θ under the alternative. Finally, Bayes factors and relative belief ratios can also be seen as special cases of 'Bayesian evidence values' whenever the Savage-Dickey density ratio applies, so interval estimates based on Bayesian evidence values also correspond to support intervals in these cases (Kelter, 2022).

4 Applications

We will now illustrate SCs with the analysis of a binomial proportion (Section 4.1), meta-analysis (Section 4.2), replication studies (Section 4.3), and logistic regression (Section 4.4).

4.1 Binomial proportion

Bartoš et al. (2025) 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 $y = 178'079$ 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. (2025) 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 $\text{B}(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1} \,\mathrm{d}t$ and the incomplete regularized beta function $I_x(a, b) = \{\int_0^x t^{a-1} (1 - t)^{b-1} \,\mathrm{d}t\} / \text{B}(a, b)$. Specifically, Bartoš et al. (2025) 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 shows the resulting SC for a range of probabilities from 0.5 to 0.515. Looking at the SC evaluated at $\theta = 0.5$, we can see the finding reported by Bartoš et al. (2025): There is extreme evidence ($\text{BF}_{01} = 1/(1.76 \times 10^{17})$) against $\theta = 0.5$ and in favour of the alternative concentrated around $\theta = 0.51$. This result hence provides decisive evidence for the theory from Diaconis et al. (2007) over the hypothesis that coins tend to land on the same side with equal probability. However, the SC 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.

Both the $k = 1$ support interval and the MEE coincide with the 95% central credible interval and posterior mean based on a uniform prior distribution which were reported by Bartoš et al. (2025)

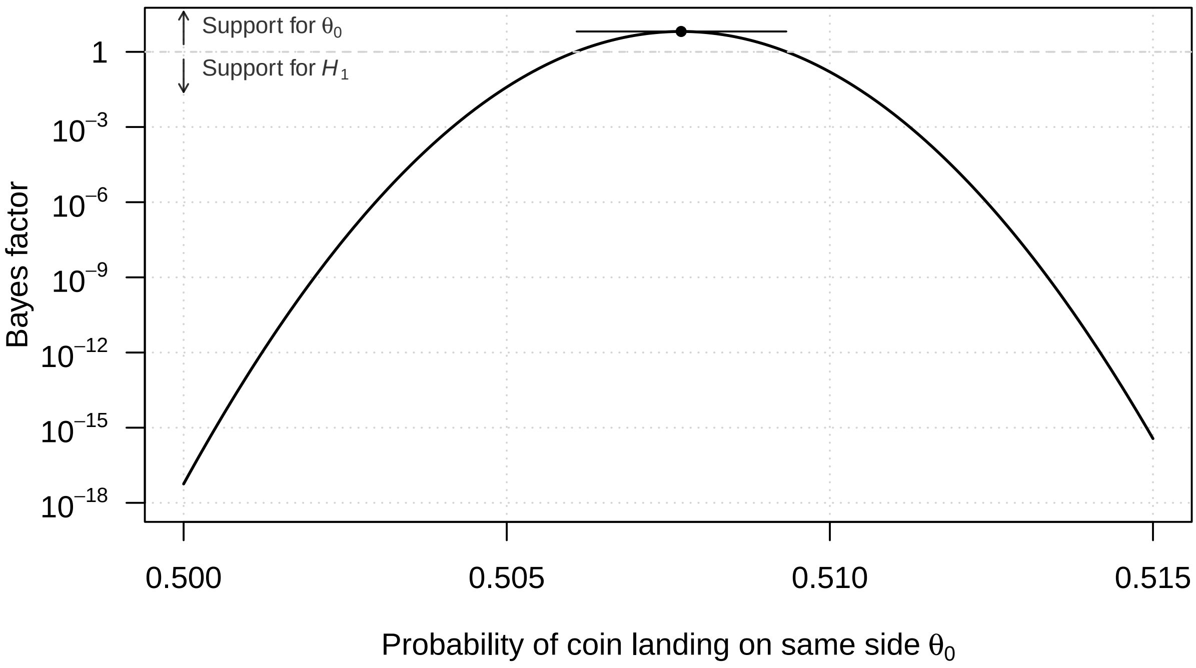

Figure 4: Support curve analysis of data from Bartoš et al. (2025). Among $n = 350'757$ coin flips, $y = 178'079$ landed on the same side as they started. A beta prior tightly concentrated around the theoretically predicted probability of 0.51 is assigned to the probability under the alternative ($\theta \mid H_1 \sim \text{Beta}(5100, 4900)_{[0.5,1]}$).

alongside the Bayes factor for $\theta = 0.5$. The difference, however, is that the MEE, Bayes factor, and support intervals are all coherently linked to the same SC based on the same prior and data model, whereas the mix of Bayes factor, posterior mean, and credible interval is not.

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 $n = 48$ 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. (2025) 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 participant specific true probability θ_i,

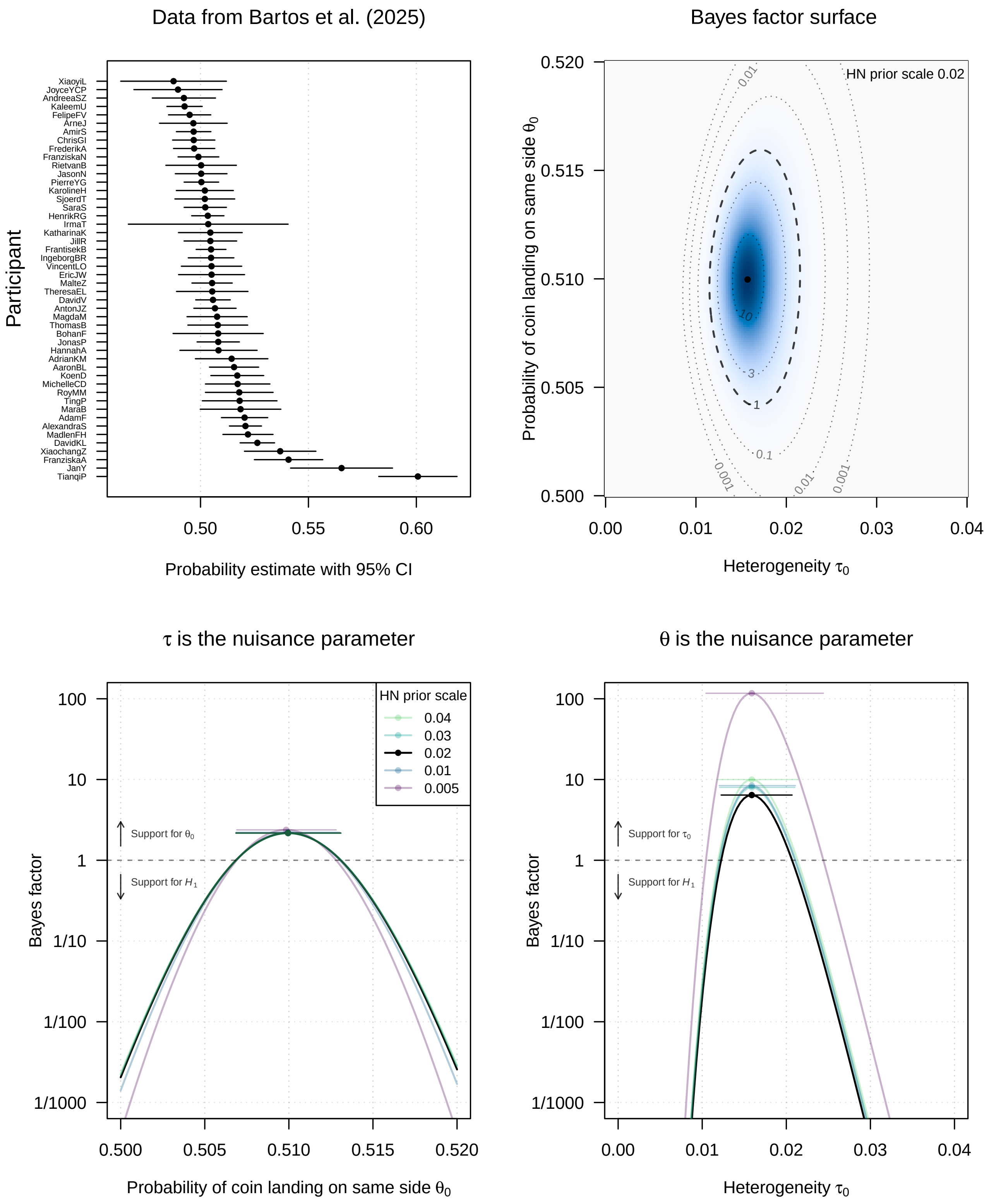

Figure 5: Bayes factor analysis of coin flipping experiments from Bartoš et al. (2025), taking into account between-participant heterogeneity. The product of a truncated beta prior ($\theta \mid H_1 \sim \text{Beta}(5100, 4900)_{[0.5,1]}$) for θ and a half-normal prior with scale 0.02 for τ are assigned 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 SC for other scale parameters of the half-normal prior.

i.e.,

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

While in some applications, inferences about the participant specific probabilities θ_i would also be of interest, these are consider nuisance parameters in this analysis. Consequently, the distribution of an estimate is marginalized over the participant probabilities distribution, leading to the model

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

There are two unknown parameters left, θ and τ. The mean θ quantifies the average probability across participants, while the heterogeneity standard deviation τ quantifies the heterogeneity of these probabilities. 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

$$\mathrm{BF}_{01}(y_1, \dots, y_n; \theta_0, \tau_0) = \frac{\prod_i^n \mathrm{N}(y_i \mid \theta_0, \sigma_i^2 + \tau_0^2)}{\int_0^\infty \int_0^1 \prod_i^n \mathrm{N}(y_i \mid \theta, \sigma_i^2 + \tau^2)\, p(\theta, \tau \mid H_1)\, \mathrm{d}\theta\, \mathrm{d}\tau}$$

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 $\theta \mid H_1 \sim \text{Beta}(5100, 4900)_{[0.5,1]}$ 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 95% probability to τ values smaller than 0.04, thus encoding the possibility of no heterogeneity (all true participant probabilities are the same when $\tau = 0$) up to small amounts of heterogeneity (the true participant probabilities differ by a few percentage points). SCs 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 SC 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 deviation ($\hat{\tau}_{\mathrm{ME}} = 0.016$) suggests small but non-negligible heterogeneity. This MEE receives strong support over the alternative ($k_{\mathrm{ME}} = 14$). 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 SC shows that probabilities of $\theta = 0.5$ and no heterogeneity $\tau = 0$ are decisively refuted by the data over the alternative ($\log \mathrm{BF}_{01} = -1.81 \times 10^5$).

The two bottom plots in Figure 5 show SCs 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, SCs 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 marginal MEEs ($\hat{\theta}_{\mathrm{ME}} = 0.51$ and $\hat{\tau}_{\mathrm{ME}} = 0.016$) align with the joint MEEs, but their evidence values ($k_{\mathrm{ME}} = 2.2$ and $k_{\mathrm{ME}} = 6.4$, respectively) indicate less support over the alternative than for the joint one. Finally, looking at the colored SCs 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. For the latter parameter, increasing the scale of the prior does not seem to change the SC too much, while decreasing the scale to a value of $s = 0.005$ dramatically increases the height of the SC, 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.

In sum, the informative hypothesis tests carried out by Bartoš et al. (2025) could be embedded in a SC framework that additionally supplies them with compatible point and interval estimates. This analysis suggests that, on average, coins tend to land on the same side with probability $\theta = 0.51$ in accordance with the hypothesis from Diaconis et al. (2007). At the same time, there seems to be non-negligible between-flipper heterogeneity, with a betwen-flipper standard deviation of around one to two percent.

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., 2023). 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 SCs.

Suppose that original and replication study provide effect estimates y_o and y_r with standard errors σ_o and σ_r, respectively. Each is supposed to be normally distributed around a common underlying effect size θ with (assumed to be known) variance equal to its squared standard error, i.e., $y_i \mid \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ for $i \in \{o, r\}$. A 'replication SC' may then be obtained by using the replication data to contrast the null hypothesis $H_0\colon \theta = \theta_0$ to the alternative $H_1\colon \theta \sim \mathrm{N}(y_o, \sigma_o^2)$, 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). As such, the replication Bayes factor represents a special case of the 'partial Bayes factor' (O'Hagan and Forster, 2004, p.186). This leads to the following SC

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

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

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

and k support interval

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

We will now demonstrate application of the replication SC by reanalyzing data from the replication project by Wagenmakers et al. (2016). This project attempted to replicate the original study from Strack et al. (1988) across 17 different study sites. The original study tested the so-called 'facial feedback hypothesis' and found that participants gave higher funniness ratings to cartoons if they were smiling as opposed to showing discontent (estimated mean difference of 0.82 units on a 10-point Likert scale, with 95% confidence interval from −0.05 to 1.69). The replications were conducted across 17 different study sites, each producing a mean difference effect estimate and confidence interval.

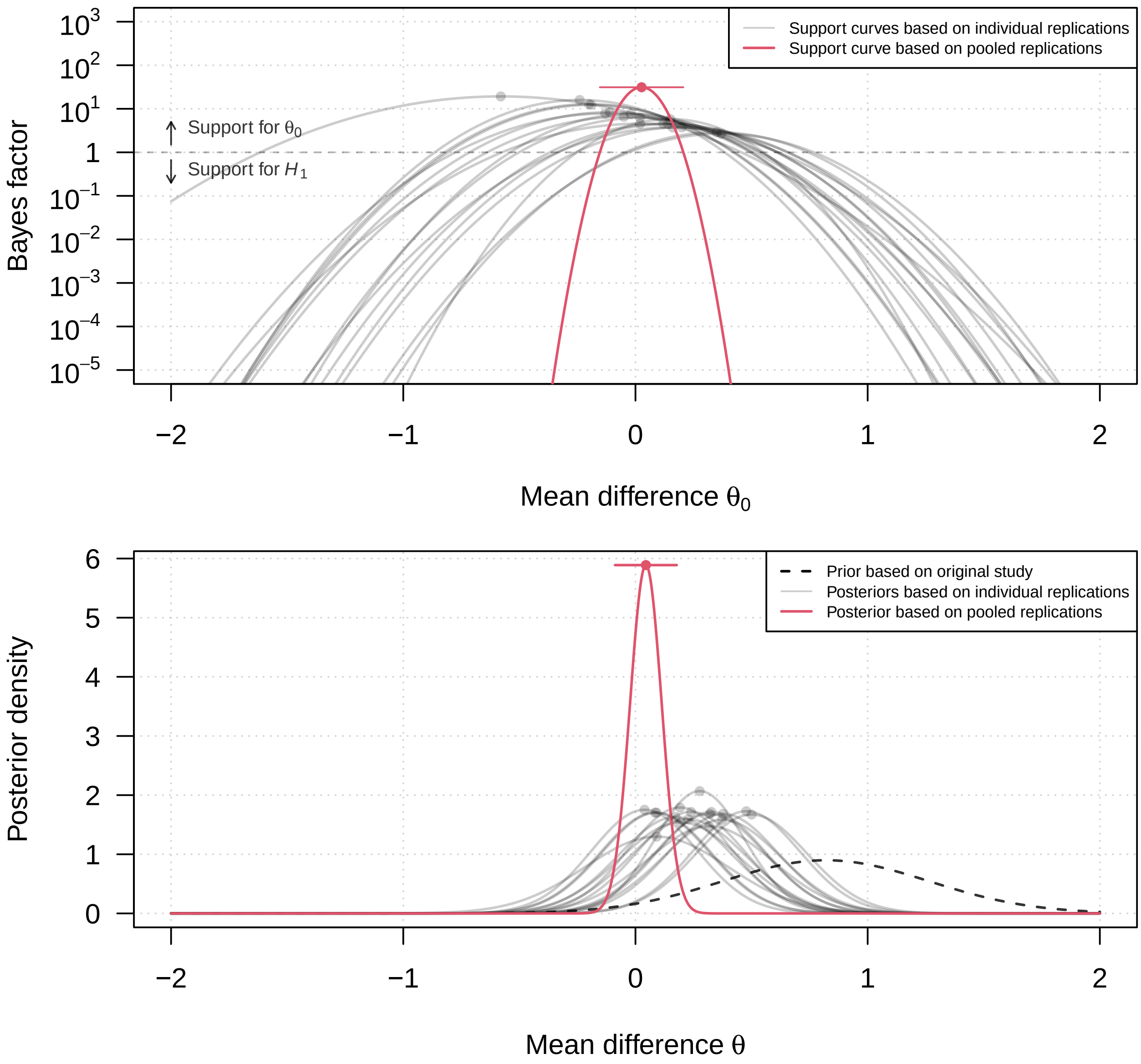

Figure 6: Support curves with maximum evidence estimate and $k = 1$ support interval (top) and posterior distribution with posterior mode and 95% highest posterior density credible interval (bottom) for the pooled replication studies from the 'facial feedback hypothesis' (Strack et al., 1988; Wagenmakers et al., 2016). The original study found an estimated standardized mean difference of $y_o = 0.82$ with standard error $\sigma_o = 0.44$ 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 are pooled via inverse variance weighting. The posterior is obtained by multiplying the SC with the prior density.

In contrast to the original study, the pooled replication mean difference was very close to zero (estimated mean difference of 0.03 with 95% confidence interval from -0.11 to 0.16). We will now assess the replicability of the original finding using replication Bayes factors and the corresponding SCs.

The top plot in Figure 6 shows the associated replication SCs, MEE, and $k = 1$ support interval. We see from the SC based on the pooled replications that mean difference values close to zero receive more evidence compared to the prior based on the original study ($\hat{\theta}_{\text{ME}} = 0.03$ with $k = 1$ support interval from -0.15 to 0.21), all of them being much smaller than the estimated mean difference from the original study 0.82. Furthermore, the SC at the mean difference of zero indicates strong evidence in favour of no difference over the prior based on the original study ($\text{BF}_{01} = 29$). Thus, the replication SC analysis suggests hardly any replicability of the original facial feedback effect.

The bottom plot in Figure 6 illustrates the posterior distributions, conveniently obtained by multiplying the SC by the prior distribution based on the original data since this SC has a Savage-Dickey density ratio representation. We can see that the $k = 1$ support intervals from the top plot are given by the set of effect sizes with posterior density larger than the prior density. We can also see that for the pooled replications the SC inferences mostly align with the posterior inferences, that is, the maximum a posteriori estimate and 95% credible interval are very close to the MEE and $k = 1$ support interval. This is because when pooling all the replication data, the data are much more informative than the prior based on the original study and thus overturn it. However, when comparing the individual replications' SCs and posterior distributions we can see much larger differences. The SCs peak at the corresponding replication effect estimates while the posteriors peak at weighted averages of original and replication estimates. For example, the replication with the smallest effect has a SC with $\hat{\theta}_{\text{ME}} = -0.58$ and with $k = 1$ support interval from -1.62 to 0.46, whereas the corresponding posterior has its mode at $\hat{\theta}_{\text{MAP}} = 0.09$ with 95% credible interval from -0.51 to 0.69. This apparent discrepancy reflects the fact that SCs do not synthesize the data with the prior, as opposed to the posterior analysis. Since in a replication study the interest is often to challenge the original study, synthesis with the original data is arguably undesirable in the analysis of replication studies. As such, a replication SC analysis, as presented here, can be useful because it can clearly show whether or not there are parameter values that are better supported than the prior based on the original study.

4.4 Logistic regression

To illustrate a more computationally demanding application of SCs, we consider the criminological study by Ejbye-Ernst et al. (2023). This study evaluated the effect of a text-based light projection intervention (*"It's illegal to buy drugs from street dealers"* and a pictogram illustrating the same message) on the presence of street drug dealers in Amsterdam's red-light district. To this end, the researchers analyzed 765 one-minute segments of video footage from two surveillance cameras installed in the red-light district with respect to the presence/absence of street dealers before and after the intervention, as well as other covariates (day of the week, camera, number of people, and hours after 8 p.m.). Out of 765 segments, street dealers were identified in 60. This small number of events makes it challenging to draw conclusive inferences about the effect of the intervention, as we will see in the following.

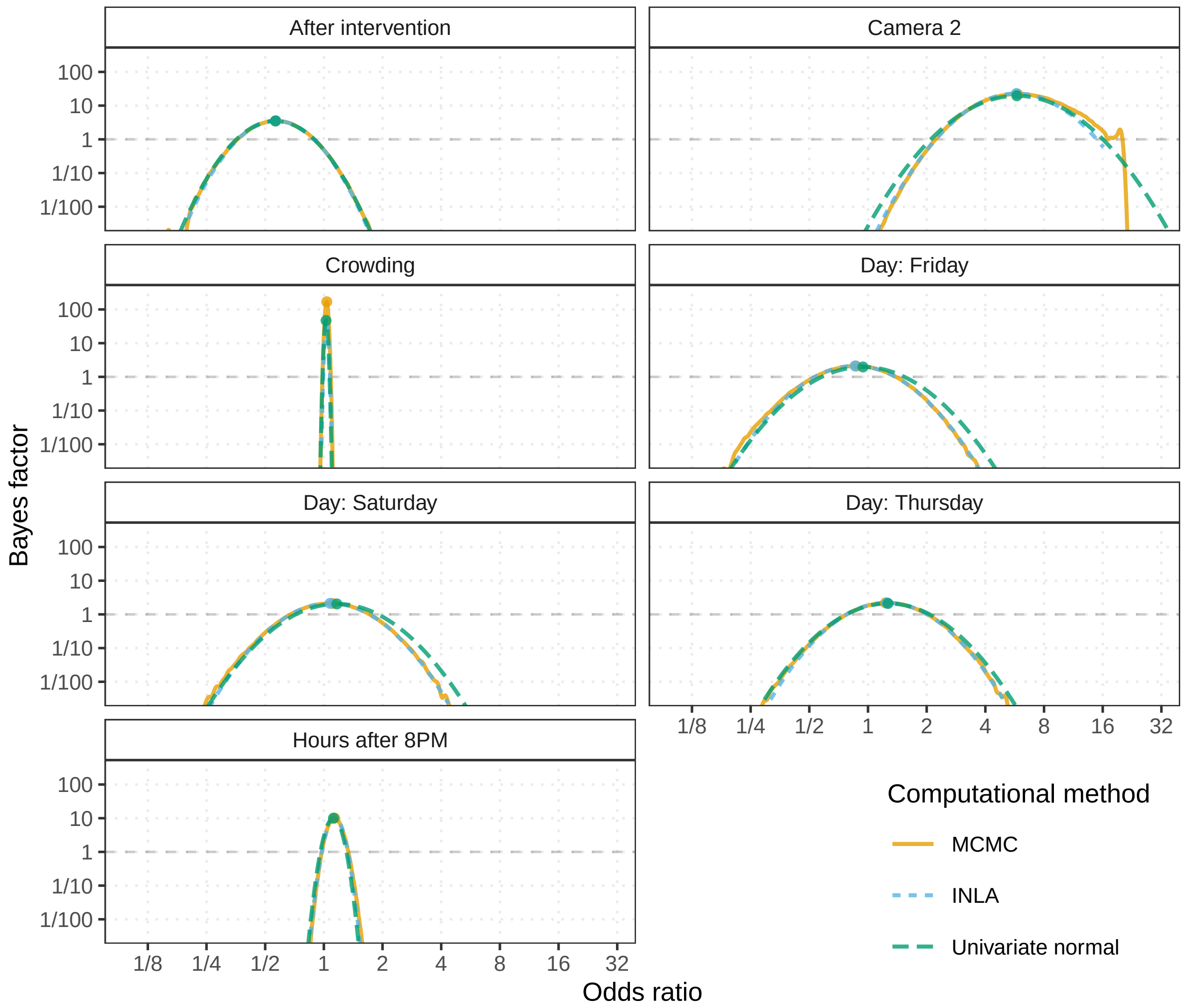

Figure 7: Multiple logistic regression SC analysis of presence of street drug dealers in the Amsterdam red-light district before and after intervention (Ejbye-Ernst et al., 2023). Each plot shows the SC related to the exponentiated regression coefficient which can be interpreted as odds ratio. Independent, weakly informative $\mathrm{N}(0, 1/2)$ priors are assigned to the coefficients under the alternative H_1. All other coefficients were considered as nuisance parameters and the same priors assigned to them as under the alternative. Variables are binary indicators, except 'Crowding' (the number of people at the beginning of a video observation) and 'Hours after 8PM'. Wednesday is the reference category for the 'Day' covariate.

Figure 7 shows the SCs related to a logistic regression analysis of the data including an intercept term and a main effect for each covariate. Each SC relates to the exponentiated regression coefficient, which can be interpreted as the multiplicative change of odds of street dealer presence 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 $\mathrm{N}(0, 1/2)$ priors were assigned to the coefficients under the alternative. This prior represents a weakly informative alternative postulating that the median odds ratio is 1 and that 95% of odds ratios are in between 1/4 and 4, representing a vague but realistic range of odds ratios in observational data (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. Note that even though we only look at the SC of a single parameter, this approach does not ignore the nuisance parameters but averages the marginal likelihood over their prior, thereby incorporating nuisance parameter uncertainty.

SCs were computed in three ways: i) By first computing the marginal posterior distribution for each coefficient from kernel smoothing of 1′000′000 Markov chain Monte Carlo (MCMC) samples (solid orange lines) computed with `Stan` (Carpenter et al., 2017) and then computing the SC via the Savage-Dickey density ratio as explained in Section 3.4. ii) By computing the marginal posterior with integrated nested Laplace approximation (INLA; short-dashed blue lines) via `INLA` (Rue et al., 2009) and then computing the SC via the Savage-Dickey density. iii) By estimating the parameters of the logistic model first with maximum likelihood, and then using each estimated coefficient and its standard error for a univariate normal analysis similar to Section 4.3, ignoring the nuisance parameters in the SC analysis (long-dashed green lines). As a result, the MEEs from the univariate normal analysis correspond to maximum likelihood estimates, while the MEEs from the MCMC and INLA analyses correspond to integrated maximum likelihood estimates.

The MCMC analysis took the longest of the three (around 5 minutes to run), followed by the INLA analysis (about a second to run), followed by the univariate analysis (almost instantaneous). *Despite the many MCMC samples that were taken to obtain smooth SCs*, we can see that the SCs based on MCMC and INLA (with default settings) may still have inaccuracies or cannot be computed in the outer regions of the SC, since these are regions where the posterior density is close to zero. We can also see that the univariate normal SC agrees reasonably well with the MCMC and INLA SCs in most cases, with the exception of the 'Camera 2' coefficient. In this case, the MCMC and INLA SCs end abruptly around OR = 16, because no larger MCMC samples were observed or because the INLA algorithm returned a posterior density of zero. In this case, the normal approximation SC could be used cautiously as an extrapolation of the MCMC and INLA SCs.

Concering the effect of the intervention, the SC (based on MCMC) for the variable 'After intervention (top left panel) has its mode at $\widehat{\text{OR}}_{\text{ME}} = 1/1.8$ indicating a negative association between the text-based intervention and the presence of street dealers, yet this parameter value receives only moderate support over the alternative ($k_{\text{ME}} = 3.5$) and the corresponding $k = 1$ support interval spans the range from a noticeable association (OR $= 1/2.8$) up to hardly any association (OR $= 1/1.1$). This suggests some negative association between the intervention and the presence of street dealers, although the extent of the association being relatively unclear.

Among the remaining variables, the SC related to 'Camera 2' clearly suggests a higher prevalence of street dealers at the location of camera 2 compared to camera 1 ($\widehat{\text{OR}}_{\text{ME}} = 5.8$ with $k_{\text{ME}} = 22.2$ and $k = 1$ support interval from OR $= 2.2$ to OR $= 17.1$) while the 'Crowding' SC suggests a small but positive association between the number of people in a video sequence and the presence of street dealers ($\widehat{\text{OR}}_{\text{ME}} = 1.04$ with $k_{\text{ME}} = 169.3$ and $k = 1$ support interval from OR $= 1$ to OR $= 1.1$). On the other hand, the SCs related to the day of the week and hours after 8 p.m. are largely undiagnostic about whether or not the variables exhibit negative or possitive associations with the presence of street dealers, possibly due to the sparse nature of the data.

In sum, this example demonstrated how SCs can be applied to more complicated models such as logistic regression, and how they can be obtained from general-purpose programs for computing posterior distributions such as `Stan` or `INLA`.

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 support curves (SCs), 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: While a P-value function can only quantify evidence *against* parameter values (Greenland, 2023), SCs allow us to quantify evidence *in favour* of parameter values over the alternative. Moreover, if a SC diagnoses absence of evidence, data analysts can continue to collect data without worrying about multiplicity issues. Like ordinary Bayesian inference, SC inference uses the Bayesian evidence calculus, but without synthesizing data and prior. When point priors are assigned, Bayes factors become likelihood ratios, so SC inference aligns with likelihoodist inference, but when there are nuisance parameters, SCs include a natural way to eliminate them via marginalization over a prior.

Like the likelihoodist and Neyman-Pearson paradigms of statistical inference, SC inference requires the formulation of alternative hypotheses. For this reason, SCs are particularly valuable in contexts where prior data or theories are available to formulate alternative hypotheses. For example, SCs (under the name of 'K ratio') have been used by the large-scale NANOGrav collaboration to quantify the evidence for new physics theories against the established Standard Model (Afzal et al., 2023). Similarly, our application of SCs to the coin flipping experiment from Bartoš et al. (2025) or the replication studies from Wagenmakers et al. (2016) was relatively straightforward as the prior distribution for the parameters under the alternative could be specified through the a priori specified hypotheses and the data from the original study, respectively. In cases where there are no clear alternative hypotheses, data analysts may use SCs based on 'weakly informative' (Gelman, 2009) or 'default' prior distributions (e.g., unit-information priors, see Kass and Wasserman, 1995), but should acknowledge this limitation and report sensitivity analyses (e.g., SCs for different prior distributions). Another possibility is to base SC inference 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., 2024).

We have focused on situations where null models are nested within the alternative model. However, SC inference could also be applied to more general model comparison where the null is not nested in the alternative. In this case, the interpretation of the SC is more intricate as the comparison does not only involve an additional parameter but an entirely different model. Moreover, computing the SC via the Savage-Dickey density ratio is not possible anymore.

Where under their control, data analysts should design experiments and studies so that conclusive inferences can be drawn from the data collected, including SC inferences. Future research needs to investigate how experiments need to be designed to enable conclusive inference with SCs. For example, one may design an experiment so that the expected width of a support interval is sufficiently narrow, or so that the expected evidence level for the MEE is sufficiently large. Finally, calculating SCs can be challenging, as our logistic regression example showed. For example, if a SC is computed via the Savage-Dickey density ratio from a posterior distribution computed by MCMC, the SC may be imprecise at the tails of the posterior, even with millions of samples. Such problems become more pronounced in high-dimensional settings. Future work may focus on developing more efficient techniques for computing SCs 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. In fact, there is empirical evidence that researchers sometimes misuse or misinterpret Bayes factors. For example, Bayes factors are sometimes erronously interpreted as absolute rather than relative evidence, or as posterior odds rather than updating factors (Wong et al., 2022; Tendeiro et al., 2024). Bayes factors by construction do not take into account the prior probabilities of their contrasted hypotheses, so they may indicate strong support for a hypothesis even though this hypothesis would still remain unlikely when combined with its prior probability (Lavine and Schervish, 1999; Good, 2001). Any type of statistical inference can lead to distorted scientific inferences if used in a bright-line fashion without consideration of contextual factors (Goodman, 2016; Greenland, 2023). We believe that SCs 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 Ejbye-Ernst et al. (2023) and Bartoš et al. (2025) for openly sharing their data. We thank Riko Kelter, Andrew Fowlie, Eric-Jan Wagenmakers, František Bartoš, Leonhard Held, Małgorzata Roos, and Sander Greenland for valuable comments on drafts of the manuscript. We thank Yang Liu, Jonathan Williams, and another anonymous reviewer for helpful and constructive comments that substantially improved paper. 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 data from Ejbye-Ernst et al. (2023) where obtained from the '`data.RData`' file available at `https://osf.io/nb56d`. The data from Wagenmakers et al. (2016) were manually extracted from Figure 4 in their paper. The standard errors were recomputed by dividing the difference of the upper and the lower confidence interval bounds by twice the 97.5% quantile of the standard normal distribution. The data from Bartoš et al. (2025) were obtained from the `dat.bartos2023` data set included in the `metadat` R package (White et al., 2023). 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://doi.org/10.5281/zenodo.10817311`. We used the statistical programming language R version 4.4.1 (2024-06-14) for analyses (R Core Team, 2023) along with the `brms` (Bürkner, 2021) and `INLA` (Rue et al., 2009) packages for the computation of posterior distributions.

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., et al. (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., et al. (2025). Fair coins tend to land on the same side they started: Evidence from 350, 757 flips. *Journal of the American Statistical Association*. doi:10.1080/01621459.2025.2516210. in press.

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.

Bind, M.-A. C. and Rubin, D. (2024). Counternull sets in randomized experiments. *The American Statistician*, pages 1–16. doi:10.1080/00031305.2024.2432884.

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.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54. doi:10.18637/jss.v100.i05.

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):627–679. 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.

Ejbye-Ernst, P., Moeller, K., Liebst, L. S., Thomas, J., Sexton, M., and Lindegaard, M. R. (2023). "it's illegal to buy drugs from street dealers"—a video-based pre-post study of a behavioral intervention to displace dealers from an Amsterdam open-air drug market. *Journal of Experimental Criminology*. doi:10.1007/s11292-023-09602-9.

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. *The European Physical Journal C*, 84(426). doi:10.1140/epjc/s10052-024-12792-9.

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.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514. doi:10.1111/j.2517-6161.1994.tb01996.x.

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.

Good, I. J. (2001). Lavine, M., and Schervish, M. J. (1999), "Bayes factors: What they are and what they are not," The American Statistician, 53, 119–122: Comment by Good and reply. *The American Statistician*, 55(2):171–174. doi:10.1198/000313001750358680.

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. (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 even 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: Statistical Methodology*. doi:10.1093/jrsssb/qkae011.

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):961–989. doi:10.1214/20-ba1234.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4). doi:10.1214/ss/1009212519.

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 Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):129–144. doi:10.1111/j.2517-6161.1992.tb01868.x.

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.

Kelter, R. (2022). The evidence interval and the Bayesian evidence value: On a unified theory for Bayesian hypothesis testing and interval estimation. *British Journal of Mathematical and Statistical Psychology*, 75(3):550–592. doi:10.1111/bmsp.12267.

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.

Lavine, M. and Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, 53(2):119–122. doi:10.1080/00031305.1999.10474443.

Liu, C. C. and Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52(6):362–375. doi:10.1016/j.jmp.2008.03.002.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Smíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., and Wagenmakers, E.-J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, 88(2). doi:10.18637/jss.v088.i02.

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*, 43(6):1271–1289. doi:10.1002/sim.10000.

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Abril Pla, O., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2024). Prior knowledge elicitation: The past, present, and future. *Bayesian Analysis*, 19(4). doi:10.1214/23-ba1381.

Morey, R. D. and Rouder, J. N. (2024). *BayesFactor: Computation of Bayes Factors for Common Designs*. URL `https://CRAN.R-project.org/package=BayesFactor`. R package version 0.9.12-4.7.

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

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. (2023). 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. (2024). Evidential calibration of confidence intervals. *The American Statistician*, 78(1):1–11. doi:10.1080/00031305.2023.2216239.

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.

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.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2):216–232. doi:10.1086/675729.

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.

Rouder, J. N., Haaf, J. M., and Vandekerckhove, J. (2018). Bayesian inference for psychology, part iv: parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1):102–113. doi:10.3758/s13423-017-1420-7.

Rouder, J. N. and Morey, R. D. (2018). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 73(2):186–190. doi:10.1080/00031305.2017.1341334.

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.

Sanborn, A. N. and Hills, T. T. (2013). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, 21(2):283–300. doi:10.3758/s13423-013-0518-9.

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):909–922. doi:10.1214/14-ba883.

Sinharay, S. and Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56(3):196–201. doi:10.1198/000313002137.

Stefan, A. M., Evans, N. J., and Wagenmakers, E.-J. (2022). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, 27(2):177–197. doi:10.1037/met0000354.

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

Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54(5):768–777. doi:10.1037/0022-3514.54.5.768.

Tendeiro, J. N. and Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, 24(6):774–795. doi:10.1037/met0000221.

Tendeiro, J. N., Kiers, H. A. L., Hoekstra, R., Wong, T. K., and Morey, R. D. (2024). Diagnosing the misuse of the Bayes factor in applied research. *Advances in Methods and Practices in Psychological Science*, 7(1). doi:10.1177/25152459231213371.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the bayes factor. *Journal of Mathematical Psychology*, 54(6):491–498. doi:10.1016/j.jmp.2010.07.003.

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., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupianez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Ozdoğru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., and Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6):917–928. doi:10.1177/1745691616674458.

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F., and Etz, A. (2022). The support interval. *Erkenntnis*, 87(2):589–601. 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.

Wagenmakers, E.-J. and Ly, A. (2023). History and nature of the Jeffreys-Lindley paradox. *Archive for History of Exact Sciences*, 77:25–72. doi:10.1007/s00407-022-00298-3.

White, T., Noble, D., Senior, A., Hamilton, W. K., and Viechtbauer, W. (2023). *metadat: Meta-Analysis Datasets*. URL `https://github.com/wviechtb/metadat`. R package version 1.3-0.

Williams, J. P., Ommen, D. M., and Hannig, J. (2023). Generalized fiducial factor: An alternative to the Bayes factor for forensic identification of source problems. *The Annals of Applied Statistics*, 17(1). doi:10.1214/22-aoas1632.

Wong, T. K., Kiers, H., and Tendeiro, J. (2022). On the potential mismatch between the function of the Bayes factor and researchers' expectations. *Collabra: Psychology*, 8(1). doi:10.1525/collabra.36357.

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

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

## 2025-08-29 16:48:44.776149 Europe/Zurich

sessionInfo()

## R version 4.4.1 (2024-06-14)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 24.04.3 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0
##
## 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] haven_2.5.5   ggplot2_3.5.2 INLA_25.06.07 Matrix_1.7-2  brms_2.22.0
## [6] Rcpp_1.1.0    metabf_0.1    metadat_1.4-0 knitr_1.50
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.1      farver_2.1.2          dplyr_1.1.4
##  [4] loo_2.8.0             TH.data_1.1-2         tensorA_0.36.2.1
##  [7] mathjaxr_1.6-0        digest_0.6.37         estimability_1.5
## [10] lifecycle_1.0.4       sf_1.0-19             StanHeaders_2.32.7
## [13] survival_3.7-0        magrittr_2.0.3        posterior_1.6.1
## [16] compiler_4.4.1        rlang_1.1.5           tools_4.4.1
## [19] bridgesampling_1.1-2  sp_2.1-3              pkgbuild_1.4.4
## [22] classInt_0.4-10       curl_5.2.1            multcomp_1.4-25
## [25] abind_1.4-5           KernSmooth_2.23-24    withr_3.0.2
## [28] grid_4.4.1            stats4_4.4.1          xtable_1.8-4
## [31] e1071_1.7-14          colorspace_2.1-1      inline_0.3.19
## [34] emmeans_1.10.1        scales_1.3.0          MASS_7.3-64
## [37] cli_3.6.4             mvtnorm_1.3-3         generics_0.1.3
## [40] RcppParallel_5.1.11-1 DBI_1.2.3             proxy_0.4-27
## [43] rstan_2.32.6          stringr_1.5.1         splines_4.4.1
```

```
## [46] bayesplot_1.11.1      parallel_4.4.1        matrixStats_1.3.0
## [49] vctrs_0.6.5           V8_6.0.0              sandwich_3.1-0
## [52] jsonlite_1.8.8        hms_1.1.3             units_0.8-5
## [55] glue_1.8.0            codetools_0.2-19      distributional_0.4.0
## [58] stringi_1.8.4         cubature_2.1.0        gtable_0.3.6
## [61] QuickJSR_1.1.3        munsell_0.5.1         tibble_3.2.1
## [64] pillar_1.10.1         Brobdingnag_1.2-9     R6_2.6.1
## [67] fmesher_0.1.5         evaluate_0.24.0       lattice_0.22-6
## [70] highr_0.11            backports_1.5.0       MatrixModels_0.5-3
## [73] rstantools_2.4.0      class_7.3-22          coda_0.19-4.1
## [76] gridExtra_2.3         nlme_3.1-167          checkmate_2.3.1
## [79] xfun_0.52             zoo_1.8-12            forcats_1.0.0
## [82] pkgconfig_2.0.3
```