Evidential Calibration of Confidence Intervals

Samuel Pawel*, **Alexander Ly**†, **Eric-Jan Wagenmakers**†
* Department of Biostatistics, University of Zurich
† Department of Psychological Methods, University of Amsterdam
E-mail: samuel.pawel@uzh.ch

May 27, 2022

This is a preprint which has not yet been peer reviewed.

Abstract

We present a novel and easy to use method for calibrating error-rate based confidence intervals to evidence-based support intervals. Support intervals are obtained from inverting Bayes factors based on the point estimate and standard error of a parameter estimate. A k support interval can be interpreted such as "the interval contains parameter values under which the observed data are at least k times more likely than under a specified alternative hypothesis". Support intervals depend on the specification of an alternative hypothesis, and we present several types that allow data analyst to encode different forms of external knowledge. We also show how specification of an alternative hypothesis can to some extent be avoided by considering a class of prior distributions for the parameter under the alternative, and then computing so-called minimum support intervals which have a one-to-one mapping with confidence intervals. For the class of all alternatives, the minimum support interval corresponds to the likelihood ratio support interval which is the narrowest support interval possible. Other classes lead to wider and more conservative minimum support intervals that may better reflect the available evidence. We also show how the sample size of a future study can be determined based on the concept of support. Finally we show how the universal bound for the type-I error rate of Bayes factors leads to a bound for the coverage of support intervals, holding even under sequential analyses with optional stopping. An application to data from a clinical trial illustrates how the calibration to support intervals can lead to inferences that are both intuitive and informative.

Keywords: Bayes factor, Bayes factor bound, Bayesian hypothesis testing, coverage, type-I error rate, minimum Bayes factor, support interval, universal bound

1 Introduction

A pervasive problem in data analysis is to draw inferences about unknown parameters of statistical models. For instance, data analysts are often interested in identifying a set of parameter values which are relatively compatible with the observed data. Here we focus on a particular method for doing so —the *support set*— that arguably represents a natural evidential answer to the problem (Edwards, 1971; Royall, 1997). The evidential paradigm defines statistical evidence via the *Law of Likelihood* (Hacking, 1965), that is, data constitute evidence for one hypothesis over an alternative hypothesis if the likelihood of the data under that hypothesis is larger than

under the alternative. The likelihood ratio (or Bayes factor) measures the strength of evidence, and it plays also a central role in the construction of support sets, as we will explain in the following.

Let $f(x \mid \theta)$ denote the density (or probability mass) function of the observed data x. Let θ be an unknown parameter and denote by

$$\mathrm{BF}_{01}(x; \theta_0) = \frac{f(x \mid \mathcal{H}_0)}{f(x \mid \mathcal{H}_1)} = \frac{f(x \mid \theta_0)}{\int f(x \mid \theta)\, f(\theta \mid \mathcal{H}_1)\, \mathrm{d}\theta} \tag{1}$$

the Bayes factor quantifying the strength of evidence which the observed data x provide for the simple null hypothesis $\mathcal{H}_0 \colon \theta = \theta_0$ relative to a (possibly composite) alternative hypothesis $\mathcal{H}_1 \colon \theta \neq \theta_0$, with $f(x \mid \mathcal{H}_1)$ the marginal density of x obtained from integrating the density $f(x \mid \theta)$ with respect to the prior density of the parameter $f(\theta \mid \mathcal{H}_1)$ under the alternative $\mathcal{H}_1$ (Jeffreys, 1961; Kass and Raftery, 1995). A k *support set* for θ is then given by

$$\mathrm{S}_k = \left\{\theta_0 : \mathrm{BF}_{01}(x; \theta_0) > k\right\}, \tag{2}$$

that is, all parameter values for which the data are k times more likely than under the alternative hypothesis $\mathcal{H}_1$ (Wagenmakers et al., 2020). A k support set thus represents parameter values for which the data provide statistical evidence of at least level k.

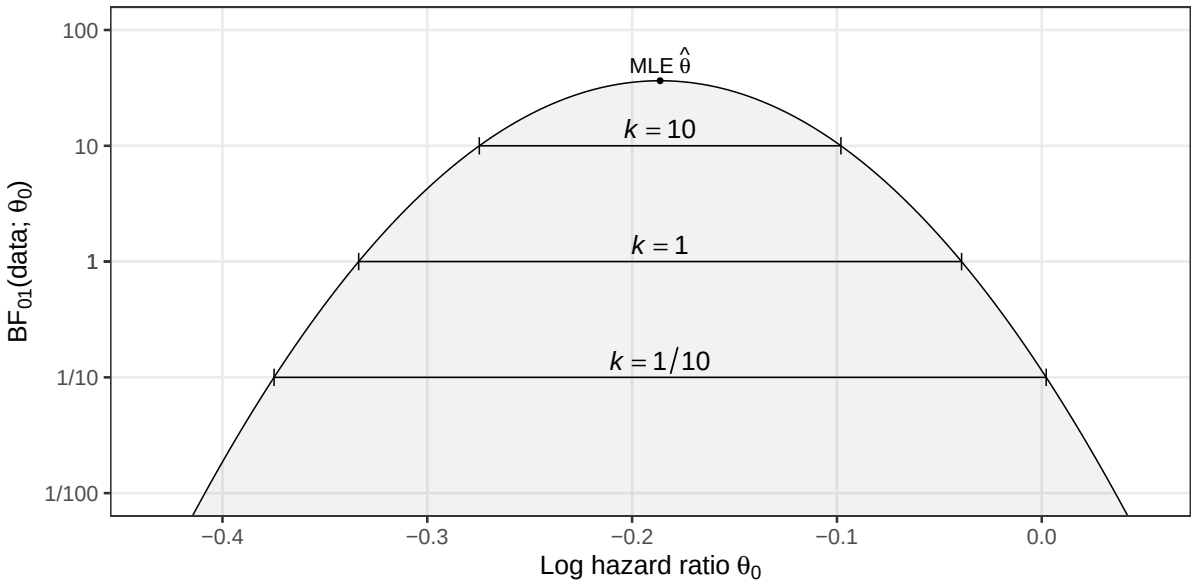

Figure 1: Application of Bayes factor functions and support intervals to data from the RECOVERY trial (RECOVERY Collaborative Group, 2021). The trial led to an age-adjusted log hazard ratio of $\hat{\theta} = -0.19$ (95% confidence interval from -0.29 to -0.07) for Covid-19 mortality in hospitalized patients treated with dexamethasone compared to usual care. The Bayes factor for testing $\mathcal{H}_0 \colon \theta = \theta_0$ versus $\mathcal{H}_0 \colon \theta \neq \theta_0$ is shown as a function of the null value θ_0. A normal distribution centered around the clinically relevant proportional mortality reduction $m = \log(0.8) = -0.22$ (as deemed by the trial steering committee) with unit-standard deviation $s = 2$ (Spiegelhalter et al., 2004, Ch. 2.4.2) is used as a prior for θ under the alternative $\mathcal{H}_1$. Different k support intervals are shown.

Figure 1 illustrates the construction of a support set based on the (soon to be discussed) data from a clinical trial. Shown is the *Bayes factor function*, i. e., the Bayes factor (1) as a function of the null value θ_0. A k support set is then given by the parameter values for which the Bayes

factor function is larger than k. For instance, the $k = 10$ support set contains parameter values that are strongly supported by the data, whereas the $k = 1/10$ support set contains values that are at least not strongly contradicted.

The construction of support sets thus parallels the construction of frequentist confidence sets: A $(1-\alpha)100\%$ confidence set corresponds to the set of parameter values which are not rejected by a null hypothesis significance test at level α. It can equally be displayed and obtained from a so-called *p-value function* (Fraser, 2019; Rafi and Greenland, 2020). Despite these similarities, the interpretation of support and confidence sets is rather different; support sets contain parameter values for which there is at least a certain amount of statistical evidence, whereas confidence sets are defined through the long-run property of including the unknown parameter θ with probability equal to their confidence level. While confidence sets are frequently interpreted to contain parameter values "compatible" with the data, their definition is based on error-rates and their confidence level does not directly relate to the strength of compatibility.

In this article we shed light on the connection between support and confidence sets. Specifically, we provide methods for calibrating approximate confidence sets to approximate support sets and vice versa in the important case when the data consists of the maximum likelihood estimate (MLE) of a univariate parameter θ (Section 2). Our main results are easy to use formulas for computing support intervals that only require summary statistics typically reported in research articles, e. g., point estimates, standard errors, or confidence intervals. Calibrating confidence to support intervals requires the specification of a prior distribution for θ under the alternative $\mathcal{H}_1$, and we compare several classes of distributions. We also show how bounding the evidence against the null hypothesis for a certain class of prior distributions leads to so-called *minimum support sets* which have a one-to-one mapping with confidence sets and thus give them an evidential interpretation (Section 3). We then illustrate how the sample size of a future study can be determined based on support, which provides an alternative to the conventional approaches based on either power or precision (Section 4). Finally, we show how the universal bound for the type-I error rate of Bayes factors can be used for bounding the coverage of support sets, even under sequential analyses with optional stopping (Section 5). To illustrate the methodology, we use data from the RECOVERY trial (RECOVERY Collaborative Group, 2021) as a running example.

2 Support intervals under normality

Denote by $\hat{\theta}$ the MLE for a univariate unknown parameter θ based on the observed data, and let σ be the (assumed to be known) standard error of $\hat{\theta}$. Assume that the usual regularity conditions for construction of a Wald-type confidence interval for θ are satisfied, so that an approximate normal likelihood $\hat{\theta} \,|\, \theta \sim \mathrm{N}(\theta, \sigma^2)$ is justifiable. In this case, an approximate $(1 - \alpha)100\%$ confidence interval for θ is given by

$$\hat{\theta} \pm \sigma \times \Phi^{-1}(1 - \alpha/2) \tag{3}$$

with $\Phi^{-1}(\cdot)$ the quantile function of the standard normal distribution. The confidence level $(1-\alpha)100\%$ represents the long run frequency with which the true parameter is included in the confidence interval (assuming that the sampling model is correct). Note that the interval (3) also corresponds to the $(1-\alpha)$ posterior credible interval based on an (improper) uniform prior for θ,

thus also representing the default interval estimate for θ from a Bayesian estimation perspective. We will now contrast the confidence interval (3) to several types of support intervals.

2.1 Normal prior under the alternative

To construct a support interval for θ, specification of a prior for θ under the alternative $\mathcal{H}_1$ is required. Specifying a normal prior $\theta \mid \mathcal{H}_1 \sim \mathrm{N}(m, s^2)$ results in the Bayes factor

$$\mathrm{BF}_{01}(\hat{\theta}; \theta_0) = \sqrt{1 + s^2/\sigma^2} \exp\left[-\frac{1}{2}\left\{\frac{(\hat{\theta} - \hat{\theta}_0)^2}{\sigma^2} - \frac{(\hat{\theta} - m)^2}{\sigma^2 + s^2}\right\}\right]. \tag{4}$$

Now, fixing the Bayes factor (4) to k and solving for θ_0 leads to the k support interval

$$\hat{\theta} \pm \sigma \times \sqrt{\log\left(1 + s^2/\sigma^2\right) + \frac{(\hat{\theta} - m)^2}{\sigma^2 + s^2} - 2\log k}. \tag{5}$$

Similar to the confidence interval (3), the support interval (5) is centered around the MLE $\hat{\theta}$. However, while the width of the confidence interval is only determined through the confidence level $(1-\alpha)$ and standard error σ, the width of the support interval also depends on the specified prior for θ under $\mathcal{H}_1$. Moreover, for $k > 1$ it may happen that the support interval does not exist, as the term below the square root in (5) may become negative for too large $k > 1$. This means that in order to find the desired level of support $k > 1$, the data have to be sufficiently informative (relative to the prior), i. e., the standard error σ has to be sufficiently small relative to the standard deviation of the prior s. In the following, we will discuss how different prior means m and standard deviations s affect the resulting support intervals.

A point prior at the MLE ($m = \hat{\theta}$ and $s \downarrow 0$) produces the likelihood ratio support interval. This type is the narrowest support interval possible, and it only exists for $k \leq 1$ as any other parameter value $\theta \neq \hat{\theta}$ receives less support than $\theta = \hat{\theta}$. In contrast, for priors that become increasingly diffuse ($s \to \infty$), the $k \geq 1$ support interval (5) extends to the entire real number line, indicating that all values $\theta \in \mathbb{R}$ receive more support from the data than the diffuse alternative. This particular behavior provides another perspective on the well-known Jeffreys-Lindley paradox (Wagenmakers and Ly, 2021); the confidence interval from (3) only spans a finite range around the MLE $\hat{\theta}$, so that the corresponding null hypothesis significance tests would reject the parameter values outside, whereas for the same values the Bayes factor would indicate evidence for the null hypothesis. Finally, centering the prior around the MLE ($m = \hat{\theta}$) and setting the variance equal to the variance of one effective observation ($s^2 = n \times \sigma^2$ with n the effective sample size), produces the support interval for Jeffreys's approximate Bayes factor (Wagenmakers, 2022). In this case, the standard error multiplier has a particularly simple form $\mathrm{M} = \sqrt{\{\log(1+n) - 2\log k\}}$, showing that at least $n \geq k^2 - 1$ effective observations are required for the respective support interval with $k \geq 1$ to exist.

2.2 Local normal prior under the alternative

The objective Bayesian approach usually specifies *local priors* under the alternative, i. e., unimodal and symmetric priors centered around the null value θ_0, for instance, the unit-information prior $\theta \mid \mathcal{H}_1 \sim \mathrm{N}(m = \theta_0, s^2 = n \times \sigma^2)$ from Kass and Wasserman (1995). This leads to a slightly

different expression for the Bayes factor

$$\text{BF}_{01}(\hat{\theta}; \theta_0) = \sqrt{1 + s^2/\sigma^2} \exp\left\{-\frac{1}{2} \frac{(\hat{\theta} - \theta_0)^2}{\sigma^2(1 + \sigma^2/s^2)}\right\} \tag{6}$$

and the corresponding k support interval

$$\hat{\theta} \pm \sigma \times \sqrt{\{\log(1 + s^2/\sigma^2) - 2\log k\}\,(1 + \sigma^2/s^2)}. \tag{7}$$

The interpretation of the support interval based on local normal priors (7) differs from the support interval based on normal priors (5) since the prior distribution of θ under the alternative hypothesis $\mathcal{H}_1$ is different for each parameter value. As such, the likelihood of the data under the alternative represents a locally averaged likelihood for each parameter value. In the case of the unit-information prior support interval, the standard error multiplier $\text{M} = \sqrt{[\{\log(1 + n) - 2\log k\}(1 + 1/n)]}$ is wider than for the Jeffreys's approximate Bayes factor by a factor of $\sqrt{(1 + 1/n)}$ but the condition $n \geq k^2 - 1$ for the existence of the $k \geq 1$ support interval is the same.

2.3 Nonlocal normal moment prior under the alternative

An alternative philosophy in Bayesian hypothesis testing is to specify so-called *nonlocal priors* under the alternative. This class of priors is characterized by having no mass at the null-value θ_0, thereby leading to a faster accumulation of evidence than local priors when the null hypothesis is actually true (Johnson and Rossell, 2010). One popular type of nonlocal priors is given by *normal moment priors* $\theta \sim \text{NM}(m, s^2)$, which have density $f(\theta \mid m, s^2) = \text{N}(\theta \mid m, s^2) \times (\theta - m)^2/s^2$ where $\text{N}(\cdot \mid m, s^2)$ denotes the density function of a normal distribution with mean m and variance s^2. The Bayes factor contrasting $\mathcal{H}_0 \colon \theta = \theta_0$ to $\mathcal{H}_1 \colon \theta \sim \text{NM}(\theta_0, s^2)$ is then given by

$$\text{BF}_{01}(\hat{\theta}; \theta_0) = (1 + s^2/\sigma^2)^{3/2} \exp\left\{-\frac{1}{2} \frac{(\hat{\theta} - \theta_0)^2}{\sigma^2(1 + \sigma^2/s^2)}\right\} \left(1 + \frac{(\hat{\theta} - \theta_0)^2}{\sigma^2(1 + \sigma^2/s^2)}\right)^{-1}$$

from which the corresponding k support interval can be derived to be

$$\hat{\theta} \pm \sigma \times \sqrt{\left[2\text{W}_0\left\{\frac{(1 + s^2/\sigma^2)^{3/2}\sqrt{e}}{2k}\right\} - 1\right](1 + \sigma^2/s^2)} \tag{8}$$

with $\text{W}_0(\cdot)$ denoting the principal branch of the Lambert W function (Corless et al., 1996). It is possible that the support interval (8) does not exist, as for the other two types of support intervals. This happens when the Lambert W term is smaller than one so that the square root is undefined. One can see from the identity $\text{W}_0(e) = 1$ that this situations occurs when $(1 + s^2/\sigma^2)^{3/2}/(2k) < \sqrt{e}$, meaning that the standard error σ has to be sufficiently small relative to the prior scale s parameter and the support level k, so that the interval exists.

2.4 Example RECOVERY trial

We now compute support intervals for the data from the RECOVERY trial (RECOVERY Collaborative Group, 2021). The trial investigated the effect of dexamethasone on the mortality of hospitalized patients with Covid-19. This led to an age-adjusted log hazard ratio of $\hat{\theta} = -0.19$

with 95% confidence interval from -0.29 to -0.07 (see the first row in Figure 2), based on which the trial investigators concluded that dexamethasone resulted in lower Covid-19 mortality compared to usual care.

The trial steering committee determined the sample size of the trial based on an assumed clinically relevant log hazard ratio of $\log 0.8 = -0.22$. This effect size can be used to inform the normal prior under the alternative $\mathcal{H}_1$, i. e., we specify the mean $m = -0.22$ along with the unit-information variance $s^2 = 4$ for a log hazard ratio (Spiegelhalter et al., 2004, chapter 2.4.2). Likewise, we use the unit-information variance $s^2 = 4$ as the variance of the local normal prior. The scale parameter of the nonlocal moment prior s is elicited with a similar approach as Pramanik and Johnson (2022); The value $s = 0.28$ is selected so that 90% probability mass is assigned to log hazard ratios between $(\theta_0 - \log 2, \theta_0)$ and $(\theta_0, \theta_0 + \log 2)$, representing effect sizes that at most double or half the hazard relative to the null value θ_0.

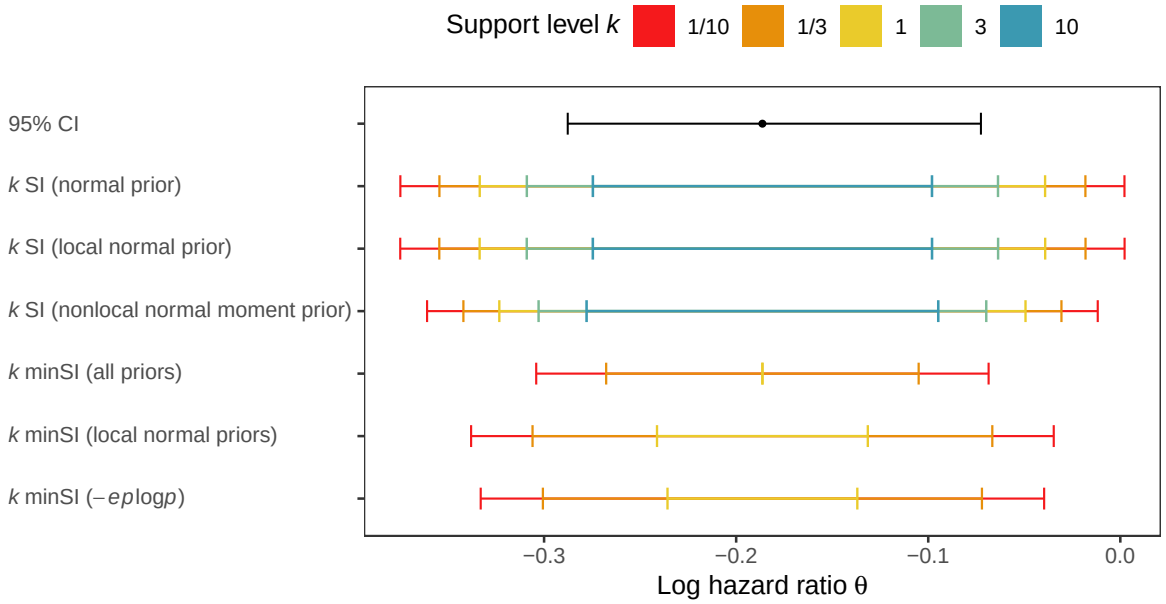

Figure 2: Comparison of different support intervals for data from RECOVERY trial. The normal prior is centered around the clinically relevant proportional mortality reduction $m = \log 0.8 = -0.22$ (as deemed by the trial steering committee) and has unit-standard deviation $s = 2$. The local normal prior also has unit-standard deviation of $s = 2$. The scale parameter of the nonlocal normal moment prior is $s = 0.28$ so that 90% of the probability is assigned to log hazard ratios between $(\theta_0 - \log 2, \theta_0)$ and $(\theta_0, \theta_0 + \log 2)$, representing effect sizes that at most double or half the mortality hazards relative to the null value θ_0.

Figure 2 shows the corresponding k support intervals for different values of k. The support intervals based on normal (second row) and local normal prior (third row) mostly coincide for all considered support levels k. The $k = 10$ support intervals (blue) from both types indicate that log hazard ratios between -0.27 and -0.1 receive strong support from the data compared to alternative parameter values. In contrast, the $k = 10$ support interval (blue) based on the nonlocal normal moment prior (fourth row) is slightly wider, indicating that values between -0.28 and -0.09 are strongly supported by the data. For smaller support levels ($k < 10$) this trend reverses and the normal and local normal prior support intervals are wider than the one based on the nonlocal normal prior. Finally, by assessing whether a k support interval includes a certain null value, one can quickly check whether the respective Bayes factor is smaller than k, similarly to the relationship between confidence intervals and p-values. For instance, one can

immediately see that the Bayes factor based on nonlocal moment priors indicates strong evidence ($\text{BF}_{01} < 1/10$) against $\mathcal{H}_0 \colon \theta = 0$ as the value is not included in the interval, whereas this is not the case for the Bayes factors based on normal and local normal priors.

3 Support intervals based on Bayes factor bounds

In some situations it is clear which prior for θ should be chosen under the alternative $\mathcal{H}_1$, e. g., when an MLE from a previous data set is available. In other situations it is less clear and different priors may produce drastically different results. To provide a more objective assessment of evidence in the latter situation, several authors have proposed to instead specify only a class of prior distributions and then select the one prior among them that leads to the Bayes factor providing the most possible evidence against the null hypothesis $\mathcal{H}_0$ (Edwards et al., 1963; Berger and Sellke, 1987; Sellke et al., 2001; Held and Ott, 2018). Depending on whether the Bayes factor is oriented in favor of the null $\mathcal{H}_0$ (as here) or the alternative $\mathcal{H}_1$, such Bayes factor bounds are called *minimum Bayes factors* or *maximum Bayes factors*, respectively.

We will now show how minimum Bayes factors can be used for obtaining so-called *minimum support sets*. Specifically, a k minimum support set is given by

$$\text{minS}_k = \{\theta_0 : \text{minBF}_{01}(x; \theta_0) \geq k\}, \tag{9}$$

where $\text{minBF}_{01}(x; \theta_0)$ is the smallest possible Bayes factor for testing $\mathcal{H}_0 \colon \theta = \theta_0$ versus $\mathcal{H}_1 \colon \theta \neq \theta_0$ that can be obtained from a class of prior distributions for θ under the alternative $\mathcal{H}_1$. That is, for each θ_0 the prior for θ under $\mathcal{H}_1$ is cherry-picked from a class of priors to obtain the lowest evidence for $\mathcal{H}_0 \colon \theta = \theta_0$ possible. Minimum support intervals thus provide a Bayes/non-Bayes compromise (Good, 1992) as they do not require specification of a specific prior distribution but still allow for an evidential interpretation of the resulting interval.

One property of minimum Bayes factors is that they can only be used to asses the maximum evidence *against* the null hypothesis but not for it. Minimum support sets inherit this property, meaning that they can only be obtained for support levels $k \leq 1$. For instance a $k = 1/3$ minimum support set includes the parameter values under which the observed data are *at most* 3 times less likely compared to under all priors from the specified class of alternative. Being unable to obtain support intervals with $k > 1$ is the price that needs to be paid for having to only specify a class of prior distributions but not a specific prior itself. We will now discuss minimum support intervals from several important classes of distributions.

3.1 Class of all distributions under the alternative

Among the class of all possible priors under $\mathcal{H}_1$, the prior which is most favorable towards the alternative is a point mass at the observed effect estimate $\mathcal{H}_1 \colon \theta = \hat{\theta}$ (Edwards et al., 1963). The resulting minimum Bayes factor is given by

$$\text{minBF}_{01}(\hat{\theta}; \theta_0) = \exp\left\{-\frac{1}{2}\frac{(\hat{\theta} - \theta_0)^2}{\sigma^2}\right\}, \tag{10}$$

which equals the standard likelihood ratio test statistic. Inverting (10) for θ_0 leads to the k minimum support interval

$$\hat{\theta} \pm \sigma \times \sqrt{-2 \log k}. \tag{11}$$

For each k, the support interval (11) is the narrowest support interval possible among all possible support intervals.

3.2 Class of local normal alternatives

When the class of priors for θ under the alternative $\mathcal{H}_1$ is given by normal distributions centered around the null value θ_0, choosing its variance to be $s^2 = \max\{(\hat{\theta} - \theta_0)^2 - \sigma^2, 0\}$ maximizes the marginal likelihood of the data under $\mathcal{H}_1$. Plugging this variance in the Bayes factor for local normal priors (6) leads to the minimum Bayes factor

$$\text{minBF}_{01}(\hat{\theta}; \theta_0) = \begin{cases} \dfrac{|\hat{\theta} - \theta_0|}{\sigma} \exp\left\{-\dfrac{(\hat{\theta} - \theta_0)^2}{2\sigma^2}\right\} \sqrt{e} & \text{if } \dfrac{|\hat{\theta} - \theta_0|}{\sigma} > 1 \\ 1 & \text{else} \end{cases} \tag{12}$$

as first shown by Edwards et al. (1963). Equating (12) to k and solving for θ_0 leads then to the k minimum support interval

$$\hat{\theta} \pm \sigma \times \sqrt{-\text{W}_{-1}(-k^2/e)}, \tag{13}$$

with $\text{W}_{-1}(\cdot)$ the branch of the Lambert W function that satisfies $\text{W}(y) \leq -1$ for $y \in [-e^{-1}, 0)$ (Corless et al., 1996). For $k = 1$, the standard error multiplier becomes $\text{M} = \sqrt{-\text{W}_{-1}(-1/e)} = 1$. Hence, all parameter values within one standard error around the observed MLE $\hat{\theta}$ are equally or more likely than under any local normal alternative.

3.3 Class of p-based alternatives

Sellke et al. (2001) proposed a minimum Bayes factor where the data are summarized through a p-value. The idea is that under the null hypothesis $\mathcal{H}_0 \colon \theta = \theta_0$, a p-value should be uniformly distributed, whereas under the alternative it should have a monotonically decreasing density characterized by the class of Beta$(\xi, 1)$ distributions (with $\xi \geq 1$). Choosing ξ such that the marginal likelihood of the data under $\mathcal{H}_1$ is maximized, leads to well-known "$-ep \log p$" minimum Bayes factor

$$\text{minBF}_{01}(p; \theta_0) = \begin{cases} -ep \log p & \text{if } p \leq e^{-1} \\ 1 & \text{else} \end{cases} \tag{14}$$

with $p = 2\{1 - \Phi(|\hat{\theta} - \theta_0|/\sigma)\}$. Equating (14) to k and solving for θ_0, leads to the k minimum support interval

$$\hat{\theta} \pm \sigma \times \Phi^{-1}\left[1 - \frac{\exp\{\text{W}_{-1}(-k/e)\}}{2}\right]. \tag{15}$$

For $k = 1$, the standard error multiplier is given by $\mathrm{M} = \Phi^{-1}[1 - \exp\{\mathrm{W}_{-1}(-1/e)\}/2] = \Phi^{-1}[1 - 1/(2e)] \approx 0.90$, so the $k = 1$ minimum support interval is just slightly tighter than the one based on local normal alternatives.

3.4 Example RECOVERY trial (continued)

The three bottom rows in Figure 2 show different types of k minimum support intervals computed for the data from the RECOVERY trial. Since minimum support intervals only exist for $k \leq 1$, only such support levels are shown. The (yellow) $k = 1$ minimum support interval for the class of all priors (fifth row) is just a point at the observed effect estimate $\hat{\theta} = -0.19$. In contrast, the (yellow) $k = 1$ minimum support intervals based on normal local priors (sixth row) and the $-ep \log p$ calibration (last row) span about one standard error around the effect estimate. Also for $k = 1/3$ (orange) and $k = 1/10$ (red), the minimum support interval based on the class of all priors is much more narrow than the ones based on local normal and $-ep \log p$, yet all of them are more narrow than the ordinary support intervals. This illustrates that minimum support intervals provide an anti-conservative assessment of support, in the same way that Bayes factor bounds provide an anti-conservative quantification of evidence.

3.5 Mapping between confidence and minimum support levels

For all types of minimum support intervals discussed so far, there is a one-to-one mapping between their support level k and the confidence level $(1 - \alpha)\%$ of the standard Wald-type confidence interval (3), see Figure 3. The conventional default level of 95% corresponds to a $k = 1/6.8$ support level for the class of all priors under the alternative, whereas it corresponds to a $k = 1/2.5$ and $k = 1/2.1$ support level for the $-ep \log p$ and local normal priors calibrations, respectively. Conversely, in order to obtain a $k = 1/10$ minimum support interval, one could take the 96.81% confidence interval for the class of all priors, whereas one could to take the 99.25% and 99.43% confidence intervals for the $-ep \log p$ and local normal priors calibrations, respectively. This mapping parallels the mapping between Bayes factor bounds and p-values (Held and Ott, 2018).

4 Design of new studies based on support

The classical way for determining the sample size of a new study is based on either (i) a specified power in a future test or (ii) a desired precision of a future confidence/credible interval. Here, we provide an alternative where the sample size of a future study is determined to have a desired level of support.

Assume we want to conduct a study and analyze the resulting MLE $\hat{\theta}$ using the support interval based on a normal prior (5). Further assume that we either specify a reasonable prior from existing knowledge or use a default prior, e. g., the prior for Jeffreys's approximate Bayes factor. The goal is now to determine the sample size n such that we can identify the parameter values which are strongly supported by the future data, for instance, with a support level $k = 10$ representing "strong" support in the classification from Jeffreys (1961). In order for the $k > 1$ support interval (5) to exist, the standard error σ of the MLE $\hat{\theta}$ needs to be sufficiently small so

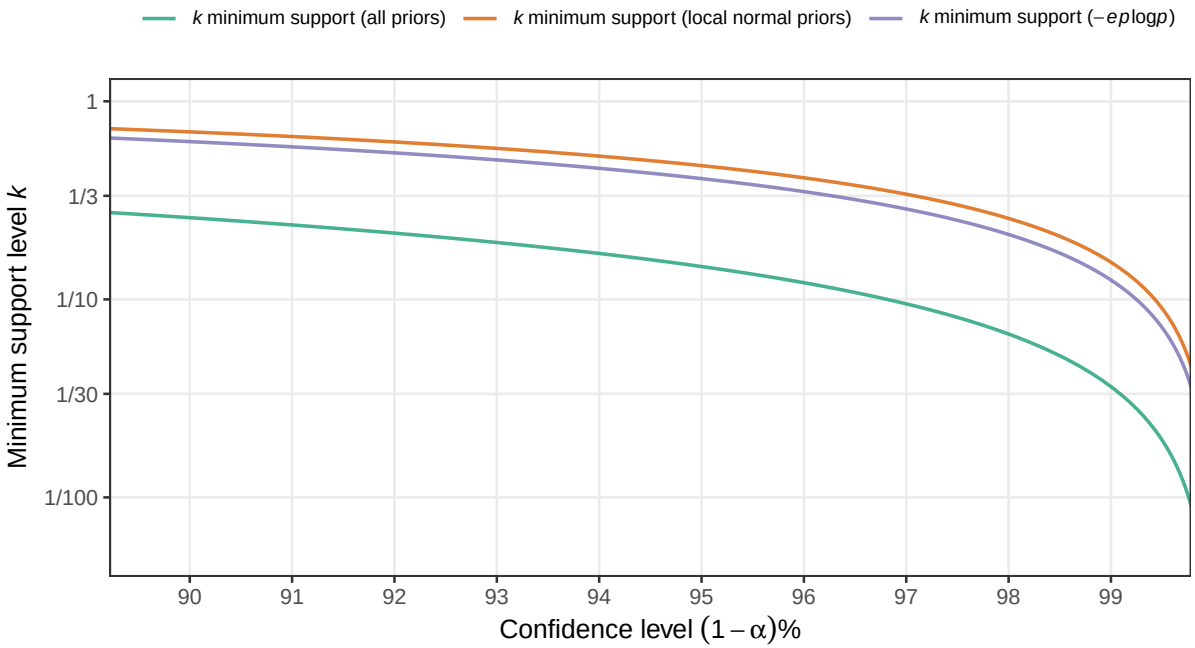

Figure 3: Mapping between confidence level $(1-\alpha)\%$ and support level k for different types of minimum support intervals.

that the term in the square root becomes non-negative, i. e., it must hold that

$$\log(1 + s^2/\sigma^2) + \frac{(\hat{\theta} - m)^2}{\sigma^2 + s^2} \geq 2 \log k. \tag{16}$$

The sample size n can now be determined such that the standard error σ is small enough for (16) to hold. The resulting sample size then guarantees that parameter values with the desired level of support will be identified. In general, this needs to be done numerically, but for the Jeffrey's approximate Bayes factor prior ($m = \hat{\theta}$ and $s^2 = n\sigma^2$), the simple expression $n \geq k^2 - 1$ mentioned earlier exists. For instance, if we want a $k = 10$ support interval to exist, we must take at least $10^2 - 1 = 99$ samples.

While the previously described approach guarantees that a $k > 1$ support interval exists for at least one parameter value θ, one may want to guarantee that the resulting k support interval will span a desired width

$$w_k = 2\sigma \times \mathrm{M}_k, \tag{17}$$

with M_k the standard error multiplier of a k support interval. In general, numerical methods are required for computing the n such that (17) is satisfied, yet again for the support interval based on Jeffrey's approximate Bayes factor there are explicit solutions available

$$n = k^2 \exp\left\{-\mathrm{W}\left(-\frac{k^2 w_k^2}{4\lambda^2}\right)\right\} \tag{18}$$

with λ^2 the variance of one (effective) observation and assuming $\log(1+n)/\log(n) \approx 1$. From (18) two things are apparent: (i) the argument to $\mathrm{W}(\cdot)$ has to be larger than $-1/e$ for the function value to be defined, meaning that the possible width is limited by $w_k \leq (4\lambda^2)/k^2$, (ii) since the

argument to $\mathrm{W}(\cdot)$ is negative, there are always two solutions given by the two real branches of the Lambert W function, if any exist at all. For instance, for a unit-information standard deviation $\lambda = 2$, a support level $k = 10$, and a desired width $w_k = 0.2$, equation (18) leads to the sample sizes $n_1 = 143$ and $n_2 = 862$ (when rounded to the next larger integer). Both lead to $k = 10$ support interval spanning the desired width, yet the Bayes factor function for the larger sample size is more peaked and can identify parameter values with more support than for the smaller sample size.

5 Error control via the universal bound

The universal bound (Appendix A) ensures that for $k \leq 1$ and when the null hypothesis $\mathcal{H}_0 \colon \theta = \theta_0$ is true, the probability for finding evidence at level k for $\mathcal{H}_0$ is smaller than k, that is

$$\Pr\left\{\mathrm{BF}_{01}(x; \theta_0) \leq k \,|\, \mathcal{H}_0\right\} \leq k \tag{19}$$

for any prior of θ under the alternative $\mathcal{H}_1$. Remarkably, the universal bound is also valid under sequential analyses and even when the data analyst actively seeks misleading evidence (Robbins, 1970). In contrast, frequentist tests and confidence sets typically have to be adjusted for sequential analyses to guarantee appropriate error rates, and the theory and applicability can become quite involved.

The universal bound can also be used to bound the coverage of a support set: Assume there is a true parameter $\theta = \theta_*$. For any alternative hypothesis $\mathcal{H}_1$, the coverage of the corresponding k support set S_k is bounded by

$$\begin{aligned} \Pr\left(\mathrm{S}_k \ni \theta_* \,|\, \theta = \theta_*\right) &= \Pr\left\{\mathrm{BF}_{01}(x \,|\, \theta_*) > k \,|\, \theta = \theta_*\right\} \\ &= 1 - \Pr\left\{\mathrm{BF}_{01}(x \,|\, \theta_*) \leq k \,|\, \theta = \theta_*\right\} \\ &\geq 1 - k \end{aligned} \tag{20}$$

where the first equality follows from the definition of a k support set (2), whereas the inequality follows from the universal bound (19). This means that any k support set with $k \leq 1$ is also a valid $(1 - k)100\%$ confidence set (provided the likelihood for the data is correctly specified).

For the case of a univariate parameter θ as considered earlier, construction of $(1 - k)\%$ approximate confidence interval via the the normal prior support interval from (5) corresponds to the proposal by Pace and Salvan (2019). These authors studied this particular case in detail and gave also frequentist motivations for the prior distributions interpreting them as weighting functions. Moreover, they found that the method also is applicable to estimates from marginal, conditional, and profile likelihoods, and that the coverage of the intervals is controlled even under slight model misspecifications. We refer to Pace and Salvan (2019) for further details.

A $(1 - k)\%$ confidence interval obtained from a k support interval will be wider than the corresponding (unadjusted) $(1 - k)\%$ Wald-type confidence interval. This is the price that has to be paid for more flexible error control. We are not claiming that support intervals are more efficient in controlling error rates than other types of adjustment methods, but they are practically much easier to implement; support intervals do not require specification of the number of interim analyses beforehand (e. g., as for Pocock or O'Brien Fleming corrections) or the recalculation of the alpha level when an interim analysis is performed (e. g., as for alpha spending functions).

This makes support intervals an appropriate tool when evidence quantification is the highest priority for data analysts, with error control as a secondary objective.

Finally, it must be noted that the coverage bound (20) only holds for support intervals but not for minimum support intervals. This is because of the one-to-one correspondence of minimum support intervals and confidence intervals. Since the coverage of confidence intervals is not controlled when repeated looks at accumulating data are performed, the same also holds true for minimum support intervals. Minimum support intervals are thus only useful for giving confidence intervals an evidential interpretation, but not for obtaining always valid confidence intervals.

6 Discussion

Misinterpretations and misconceptions of confidence intervals are common (Hoekstra et al., 2014; Greenland et al., 2016). We showed how confidence intervals can be calibrated to support intervals which have an intuitive interpretation in terms of the evidence that the data provide for the included parameter values. We obtained easy to use formulas for different types of approximate support intervals for unknown parameters based on an estimate and standard error thereof, Table 1 summarizes our results.

Table 1: Summary of confidence/credible intervals (CI), support intervals (SI), and minimum support intervals (minSI) for an unknown parameter θ based on a MLE $\hat{\theta}$ with standard error σ. All intervals are of the form $\hat{\theta} \pm \sigma \times \text{M}$. To transform an interval from type A to type B, first subtract $\hat{\theta}$ from both limits, multiply by the ratio of the standard error multipliers $\text{M}_\text{B}/\text{M}_\text{A}$, and add again $\hat{\theta}$ to both limits. The standard error multipliers M depend on either the confidence/credible level $(1 - \alpha)$ or the support level k. For the support intervals, the standard error multipliers M additionally depend on the parameters of the prior for θ under the alternative: a mean m and a standard deviation s for the normal prior $\theta \mid \mathcal{H}_1 \sim \text{N}(m, s^2)$, only a standard deviation s for the local normal prior and the nonlocal normal moment prior (see Section 2.2 and Section 2.3). The quantile function of the standard normal distribution is denoted by $\Phi^{-1}(\cdot)$, $\text{W}_0(\cdot)$ denotes the principal branch of the Lambert W function (Corless et al., 1996), and $\text{W}_{-1}(\cdot)$ denotes the branch that satisfies $\text{W}(y) \leq 1$ for $y \in [-1/e, 0)$. The respective (minimum) support intervals do only exist for support levels k for which the standard error multipliers exist, i. e., the respective term in the square root needs to be non-negative and/or the argument for $\text{W}_{-1}(\cdot)$ needs to be in $[-1/e, 0)$. All interval types can be computed with the R package `ciCalibrate`.

Interval type	Standard error multiplier M
$(1 - \alpha)100\%$ CI	$\Phi^{-1}(1 - \alpha/2)$
k SI (normal prior)	$\sqrt{\{\log\left(1 + s^2/\sigma^2\right) + (\hat{\theta} - m)^2/(\sigma^2 + s^2) - 2\log k\}}$
k SI (local normal prior)	$\sqrt{[\{\log(1 + s^2/\sigma^2) - 2\log k\}(1 + \sigma^2/s^2)]}$
k SI (nonlocal normal moment prior)	$\sqrt{([2\text{W}_0\{(1 + s^2/\sigma^2)^{3/2}/(2ke^{-1/2})\} - 1]\{1 + \sigma^2/s^2\})}$
k minSI (all priors)	$\sqrt{(-2\log k)}$
k minSI (local normal priors)	$\sqrt{\{-\text{W}_{-1}(-k^2/e)\}}$
k minSI $(-e\,p\,\log p)$	$\Phi^{-1}[1 - \exp\{\text{W}_{-1}(-k/e)\}/2]$

Which type of support interval should data analysts use in practice? We believe the support interval based on a normal prior distribution is the most intuitive for encoding external knowledge. This type should therefore be preferably used whenever external knowledge is available. At the same time, the support interval based on a local normal prior with unit-information variance (Kass and Wasserman, 1995) seems to be a reasonable "default" choice in cases where no external knowledge is available. Finally, we believe that minimum support intervals are mostly useful for

giving confidence intervals an evidential interpretation due to the one-to-one mapping between the two.

Other approaches have been proposed for calibration of confidence intervals. For instance, Rafi and Greenland (2020) propose to rename confidence intervals to "compatibility" intervals and give their confidence level an information theoretic interpretation. For example, a 95% confidence interval contains parameter values with at most 4.3 bits refutational "surprisal". We believe that this is a step in the right direction; however, we also believe that the term "compatibility" may be misleading, as absence of surprisal does not establish compatibility. To do so, an evidential framework is necessary and specification of an alternative hypothesis is unavoidable. This fact is also illustrated through the proposed minimum support intervals which can be one-to-one mapped to bits of refutational surprisal (since both can be mapped to confidence levels); without a specific alternative hypothesis only the maximum evidence *against* the included parameter values can be quantified.

We also showed how support intervals can be used for determining approximate confidence intervals whose coverage is controlled even under sequential data analyses with repeated looks at the data. Of course, such error rate guarantees rest on the assumption that the parametric data model has been correctly specified, which in most real world applications will be violated to some extent. We do not see this as a problem for the evidential interpretation of support intervals, which is usually of more concern to data analysts. Evidential inference does not rely on a statistical model being "true" in some abstract sense. Bayes factors and support intervals simply quantify the relative predictive performance that the combination of data model and parameter distribution yield on out-of-sample data (Kass and Raftery, 1995; O'Hagan and Forster, 2004; Gneiting and Raftery, 2007). As such, they are the appropriate inferential tools for making sense out of data.

Software and data

All analyses were conducted in the R programming language version 4.2.0. The code to reproduce this manuscript is available at `https://github.com/XXXXX/XXXXX`. A snapshot of the GitHub repository at the time of writing this article is archived at `https://doi.org/10.5281/zenodo.XXXXX`.

Acknowledgments

We thank Leonhard Held for helpful comments on an earlier version of the manuscript. This work was supported in part by an NWO Vici grant (016.Vici.170.083) to EJW, and a Swiss National Science Foundation mobility grant (part of 189295) to SP.

Appendix A The universal bound

We briefly review the universal bound (Royall, 1997, Ch. 1.4) for Bayes factors. Let $k \leq 1$ and $\mathcal{X}_k = \{x : \text{BF}_{01}(x) \leq k\}$ so that the probability for observing misleading evidence $\text{BF}_{01}(x) \leq k$

when $\mathcal{H}_0$ is true is

$$\begin{aligned}\Pr\left(\mathrm{BF}_{01}(X) \leq k \,|\, \mathcal{H}_0\right) &= \int_{\mathcal{X}_k} f(x \,|\, \mathcal{H}_0)\, \mathrm{d}x \\ &\leq \int_{\mathcal{X}_k} k\, f(x \,|\, \mathcal{H}_1)\, \mathrm{d}x \\ &\leq k.\end{aligned}$$

The first inequality follows from the fact that $f(x \,|\, \mathcal{H}_0) \leq k\, f(x \,|\, \mathcal{H}_1)$ for all $x \in \mathcal{X}_k$ by definition of $\mathcal{X}_k$, while the second inequality follows from the fact that the integral of $f(x \,|\, \mathcal{H}_1)$ over $\mathcal{X}_k$ can at most be one. A generalization of this result to sequential analysis of data was first established by Robbins (1970). Appendix A1 in Pace and Salvan (2019) gives his original proof in modern notation.

References

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.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359. doi:10.1007/bf02124750.

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

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242. doi:10.1037/h0044139.

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.

Gneiting, T. and Raftery, E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the Amerian Statistical Association*, 102:359–377. doi:10.1198/016214506000001437.

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419):597–606. doi:10.1080/01621459.1992.10475256.

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. 31(4):337–350. doi:10.1007/s10654-016-0149-3.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press, New York.

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.

Hoekstra, R., Morey, R. D., Rouder, J. N., and Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. 21(5):1157–1164. doi:10.3758/s13423-013-0572-3.

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

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170. doi:10.1111/j.1467-9868.2009.00730.x.

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

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

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics, volume 2B: Bayesian Inference*. Arnold, London, UK, second edition.

Pace, L. and Salvan, A. (2019). Likelihood, replicability and robbins' confidence sequences. *International Statistical Review*, 88(3):599–615. doi:10.1111/insr.12355.

Pramanik, S. and Johnson, V. E. (2022). Efficient alternatives for bayesian hypothesis tests in psychology. *Psychological Methods*. doi:10.1037/met0000482.

Rafi, Z. and Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. 20(1). doi:10.1186/s12874-020-01105-9.

RECOVERY Collaborative Group (2021). Dexamethasone in hospitalized patients with Covid-19. *New England Journal of Medicine*, 384(8):693–704. doi:10.1056/nejmoa2021436.

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.

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. Chapman & Hall, London New York.

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

Spiegelhalter, D. J., Abrams, R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley.

Wagenmakers, E.-J. (2022). Approximate objective Bayes factors from P-values and sample size: The $3p\sqrt{n}$ rule. doi:10.31234/osf.io/egydq.

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

Wagenmakers, E.-J. and Ly, A. (2021). History and nature of the Jeffreys-Lindley paradox. URL `https://arxiv.org/abs/2111.10191`.

Computational details

```
sessionInfo()

## R version 4.2.0 (2022-04-22)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## 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
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.3.6    colorspace_2.0-3 lamW_2.1.1       ciCalibrate_0.42
## [5] knitr_1.39
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.8.3       RColorBrewer_1.1-3 pillar_1.7.0       compiler_4.2.0
##  [5] highr_0.9          tools_4.2.0        digest_0.6.29      evaluate_0.15
##  [9] lifecycle_1.0.1    tibble_3.1.7       gtable_0.3.0       pkgconfig_2.0.3
## [13] rlang_1.0.2        cli_3.3.0          DBI_1.1.2          xfun_0.31
## [17] withr_2.5.0        stringr_1.4.0      dplyr_1.0.9        generics_0.1.2
## [21] vctrs_0.4.1        grid_4.2.0         tidyselect_1.1.2   glue_1.6.2
## [25] R6_2.5.1           fansi_1.0.3        farver_2.1.0       purrr_0.3.4
## [29] magrittr_2.0.3     scales_1.2.0       ellipsis_0.3.2     assertthat_0.2.1
## [33] labeling_0.4.2     utf8_1.2.2         stringi_1.7.6      RcppParallel_5.1.5
## [37] munsell_0.5.0      crayon_1.5.1
```