Evidential Calibration of Confidence Intervals

Samuel Pawel[⋆], **Alexander Ly**[†‡], **Eric-Jan Wagenmakers**[†]
⋆ Department of Biostatistics, University of Zurich
† Psychological Methods, University of Amsterdam
‡ Machine Learning Group, Centrum Wiskunde & Informatica
E-mail: samuel.pawel@uzh.ch

May 10, 2023

This is a preprint which has not yet been peer reviewed.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 a parameter estimate and its standard error. A k support interval can be interpreted as "the observed data are at least k times more likely under the included parameter values than under a specified alternative". Support intervals depend on the specification of prior distributions for the parameter under the alternative, and we present several types that allow different forms of external knowledge to be encoded. We also show how prior specification can to some extent be avoided by considering a class of prior distributions and then computing so-called minimum support intervals which, for a given class of priors, have a one-to-one mapping with confidence intervals. We also illustrate how the sample size of a future study can be determined based on the concept of support. Finally, we show how the bound for the type I error rate of Bayes factors leads to a bound for the coverage of support intervals. An application to data from a clinical trial illustrates how support intervals can lead to inferences that are both intuitive and informative.

Keywords: Bayes factor, coverage, evidence, 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 both from a likelihoodist (Edwards, 1971; Royall, 1997; Blume, 2002) and a Bayesian (Wagenmakers et al., 2022) point of view. In either paradigm, statistical evidence may be defined via the *Law of Likelihood* (Hacking, 1965), that is, data constitute evidence for one parameter value over an alternative parameter value if the likelihood of the data under that parameter value is larger than under the alternative parameter value. 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 likelihood of the observed data x. Let θ be an unknown parameter and denote by

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

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 likelihood of x obtained from integrating the likelihood $f(x \mid \theta)$ 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). For constructing a support interval, one views the Bayes factor (1) as a function of the null value θ_0 for fixed data x. A k *support set* for θ is then given by the set of parameter values for which the data are k times more likely than under the alternative hypothesis $\mathcal{H}_1$ (Wagenmakers et al., 2022), that is,

$$\text{SI}_k = \left\{\theta_0 : \text{BF}_{01}(x; \theta_0) \geq k\right\}. \tag{2}$$

The support set thus includes the parameter values for which the observed data provide statistical evidence of at least level k.

Figure 1 illustrates different support sets (in this case intervals) for a log hazard ratio parameter θ quantifying the effect of the drug dexamethasone on the mortality of hospitalized

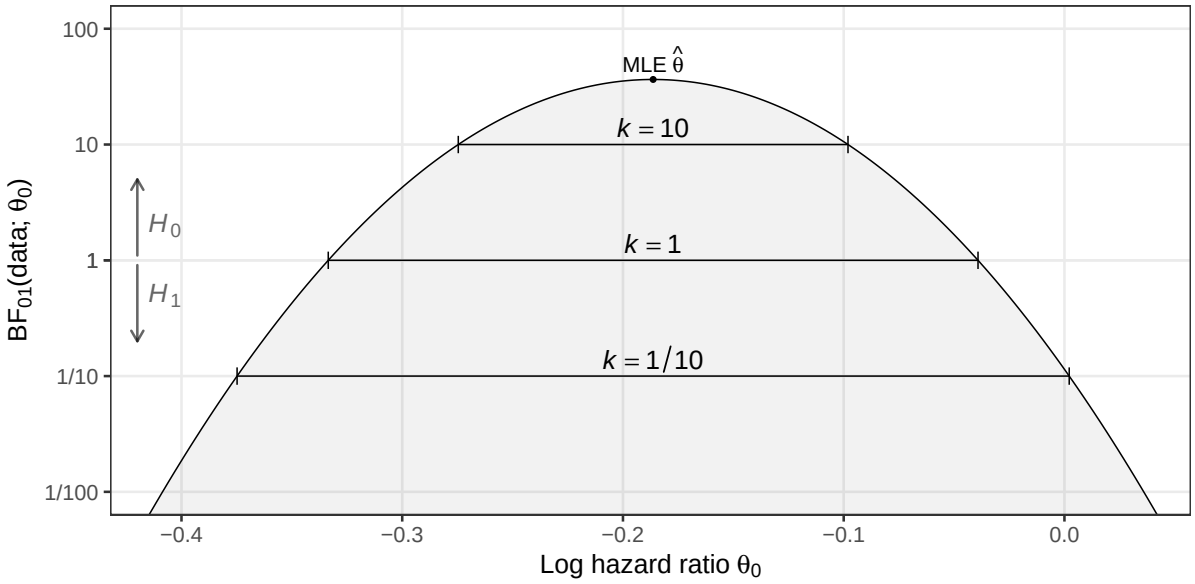


Figure 1: The RECOVERY trial (RECOVERY Collaborative Group, 2021) found that dexamethasone treatment reduced mortality compared to usual care in hospitalized Covid-19 patients (estimated log hazard ratio $\hat{\theta} = -0.19$ with standard error $\sigma = 0.05$ and 95% confidence interval from -0.29 to -0.07). Assuming a normal likelihood $\hat{\theta} \mid \theta \sim \text{N}(\theta, \sigma^2)$, the Bayes factor for contrasting $\mathcal{H}_0 \colon \theta = \theta_0$ to $\mathcal{H}_1 \colon \theta \neq \theta_0$ is shown as a function of the null value θ_0. A unit-information normal distribution $\theta \mid \mathcal{H}_1 \sim \text{N}(\mu_\theta = -0.22, \sigma^2_\theta = 4)$ centered around the clinically relevant log hazard ratio is used as prior for θ under $\mathcal{H}_1$. Support intervals for different support levels k indicate the range of log hazard ratios supported by the data.

patients with Covid-19 enrolled in the RECOVERY trial (RECOVERY Collaborative Group, 2021). Shown is also the Bayes factor for testing $\mathcal{H}_0 \colon \theta = \theta_0$ versus $\mathcal{H}_1 \colon \theta \neq \theta_0$ viewed as a function of the null value θ_0. A k support set is obtained from "cutting" this function at height k, and taking the parameter values with a Bayes factor value larger than k as part of the set. In practice, it is not clear which value of k should be chosen. One possibility is to select k based on conventional classifications of Bayes factors or likelihood ratios. Table 1 lists three of them. For instance, using the classification from Jeffreys (1961, Appendix B), the $k = 10$ support interval ranging from -0.27 to -0.1 can be interpreted to contain log hazard ratios that are *strongly supported* by the data, whereas the $k = 1/10$ support interval ranging from -0.37 to 0 can be interpreted to contain log hazard ratios that are *at least not strongly contradicted* by the data.

Table 1: Classifications of evidence for $\mathcal{H}_0$ provided by Bayes factors $\text{BF}_{01} = k$. The cut-offs from Jeffreys are slightly adjusted from powers of $\sqrt{10}$, as suggested by Held and Ott (2018). Royall and Fisher defined their classifications only for likelihood ratios, i.e., Bayes factors with simple hypotheses $\mathcal{H}_0 \colon \theta = \theta_0$ vs. $\mathcal{H}_1 \colon \theta = \theta_1$. While Royall placed no restrictions on θ_1, Fisher used the maximum likelihood estimate $\theta_1 = \hat{\theta}$. He named only the $k < 1/15$ category.

k	Jeffreys (1961)	k	Royall (1997)	k	Fisher (1956)
> 100	Decisive	> 64	Quite strong indeed	1/2 to 1	Good
30 to 100	Very strong	32 to 64	Quite strong	1/5 to 1/2	Fair
10 to 30	Strong	8 to 32	Strong	1/15 to 1/5	Poor
3 to 10	Substantial	4 to 8	Weak	$< 1/15$	Open to grave suspicion
1 to 3	Bare mention				

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*, which is the p-value of the data viewed as a function of the null value (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 frequency of including the unknown parameter θ with probability equal to their confidence level. The parameter values in a confidence sets are typically interpreted as being "compatible" with a particular data set, but this is debatable as the confidence level is concerned with the confidence set as a procedure over multiple replications.

Although support sets are conceptually simple and intuitive, they have not been applied to many problems. It is also unclear how they relate to the more widely used confidence sets. In this article we thus 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 an estimate of a univariate parameter θ with approximate normal likelihood (Section 2). To do so, we derive novel and 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. This scenario is highly relevant as a large part of commonly used estimators satisfy the approximate normality assumption, and also because one often does not have access to the raw data but

only the summary statistics. Computing a support interval 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 the novel concept of a *minimum support set*. Our minimum support sets are directly related to well-known bounds of Bayes factors (Berger and Sellke, 1987; Sellke et al., 2001; Held and Ott, 2018). In Section 3, we show how minimum support sets provide confidence sets an evidential interpretation with respect to certain classes of priors. We then illustrate how the sample size of a future study can be determined based on support, which provides a novel alternative to the conventional approaches based on either power or precision of an interval estimator (Section 5). 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 6). As a running example, we use data from the RECOVERY trial (RECOVERY Collaborative Group, 2021), as already introduced in Figure 1.

2 Support intervals under normality

Denote by $\hat{\theta}$ an asymptotically normal estimator of an unknown univariate parameter θ, possibly the maximum likelihood estimator (MLE). Suppose its squared standard error σ^2 is an estimate of the asymptotic variance of $\hat{\theta}$, so that an approximate normal likelihood $\hat{\theta} \,|\, \theta \sim \mathrm{N}(\theta, \sigma^2)$ is justifiable. For example, $\hat{\theta}$ could be an estimated regression coefficient from a generalized linear model and σ its standard error. In many simple settings, the standard error is of the form $\sigma = \lambda/\sqrt{n}$, where λ^2 is the variance corresponding to one effective unit and n is the effective sample size, for example, the number of measurements or the number of events (Spiegelhalter et al., 2004, Section 2.4), see also Berger et al. (2013) for a generalization of effective sample size to more complex settings with dependent data. 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 $(1-\alpha)100\%$ posterior credible interval based on an (improper) uniform prior for θ, corresponding to Jeffreys's transformation invariant prior (Jeffreys, 1961; Ly et al., 2017) and 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 using the data summary $\hat{\theta}$ with $\hat{\theta} \,|\, \theta \sim \mathrm{N}(\theta, \sigma^2)$, specification of a prior for θ under the alternative $\mathcal{H}_1$ is required. Specifying a normal prior $\theta \,|\, \mathcal{H}_1 \sim \mathrm{N}(\mu_\theta, \sigma_\theta^2)$

results in the Bayes factor

$$\text{BF}_{01}(\hat{\theta}; \theta_0) = \sqrt{1 + \frac{\sigma_\theta^2}{\sigma^2}} \exp\left[-\frac{1}{2}\left\{\frac{(\hat{\theta} - \theta_0)^2}{\sigma^2} - \frac{(\hat{\theta} - \mu_\theta)^2}{\sigma^2 + \sigma_\theta^2}\right\}\right]. \tag{4}$$

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

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

Similar to the confidence interval (3), the support interval (5) is centered around the parameter estimate $\hat{\theta}$. However, while the width of the confidence interval is only determined through the confidence level $(1 - \alpha)100\%$ 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 is empty, 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 squared standard error σ^2 has to be sufficiently small relative to the prior variance σ_θ^2.

In the following, we will discuss how different prior means μ_θ and variances σ_θ^2 affect the resulting support intervals. When the prior variance decreases ($\sigma_\theta^2 \downarrow 0$), the prior approaches a point mass at μ_θ. The width of the support interval is then fully determined by the difference between the parameter estimate $\hat{\theta}$ and the prior mean μ_θ divided by the standard error σ. A smaller difference between $\hat{\theta}$ and μ_θ leads to a tighter support interval. In contrast, for priors that become increasingly diffuse ($\sigma_\theta^2 \to \infty$), the $k \geq 1$ support interval (5) extends to the entire real line, indicating that all values $\theta \in \mathbb{R}$ receive more support from the data than the diffuse alternative, regardless of the data, i.e., the observed estimate $\hat{\theta}$, standard error σ, and the location of the prior mean μ_θ. This particular behavior provides another perspective on the well-known Jeffreys-Lindley paradox (Wagenmakers and Ly, 2023); the confidence interval from (3) only spans a finite range around the parameter estimate $\hat{\theta}$, 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 parameter estimate ($\mu_\theta = \hat{\theta}$) and setting the prior variance equal to the variance of one effective observation ($\sigma_\theta^2 = n \times \sigma^2$ with n the effective sample size), produces the support interval for Jeffreys's approximate Bayes factor (Wagenmakers, 2022) which is equal to well-known approximation of the Bayes factor based on the Bayesian information criterion (Raftery, 1999). In this case, the standard error multiplier has a particularly simple form $\text{M} = \sqrt{\{\log(1 + n) - 2\log k\}}$, showing that at least $n \geq k^2 - 1$ effective observations are required for the respective support interval with $k \geq 1$ to be non-empty.

2.2 Local normal prior under the alternative

The support interval based on the normal prior (5) depends on the specification of a prior mean and prior variance. A different approach is to use a so-called *local prior*, that is, a unimodal and

symmetric prior centered around the null value θ_0 (Berger and Delampady, 1987). Choosing a local normal prior with variance σ_θ^2 corresponds to setting $\mu_\theta = \theta_0$ in (4), which leads to the Bayes factor

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

The k support interval based on the Bayes factor (6) is then given by

$$\hat{\theta} \pm \sigma \times \sqrt{\left\{\log\left(1 + \frac{\sigma_\theta^2}{\sigma^2}\right) - 2\log k\right\}\left(1 + \frac{\sigma^2}{\sigma_\theta^2}\right)}. \tag{7}$$

While the Bayes factor (6) is a special case of the Bayes factor (4), the support interval (7) is not a special case of the support interval (5). This is because the prior for θ under $\mathcal{H}_1$ is different for each null value θ_0, whereas it is always the same under the two-parameter normal prior approach. To fully specify the support interval (7), the prior variance σ_θ^2 needs to be chosen. One standard choice is to set it equal to the variance of a single observation ($\sigma_\theta^2 = n \times \sigma^2$), known as unit-information prior (Kass and Wasserman, 1995). This approach leads to the k support interval

$$\hat{\theta} \pm \sigma \times \sqrt{\{\log(1 + n) - 2\log k\}(1 + 1/n)}. \tag{8}$$

For this type of 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 $k \geq 1$ support interval to be non-empty is the same.

2.3 Nonlocal normal moment prior under the alternative

Another attractive class of priors for θ under the alternative is given by so-called *nonlocal priors*. These priors are characterized by having zero density 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* $\theta \sim \text{NM}(\theta_0, \sigma_\theta)$, with symmetry point θ_0 and spread σ_θ which have density $f(\theta \,|\, \theta_0, \sigma_\theta^2) = \text{N}(\theta\,;\, \theta_0, \sigma_\theta^2) \times (\theta - \theta_0)^2/\sigma_\theta^2$ where $\text{N}(\cdot\,;\, \theta_0, \sigma_\theta^2)$ denotes the density function of a normal distribution with mean θ_0 and variance σ_θ^2. The Bayes factor employing a prior $\theta \,|\, \mathcal{H}_1 \sim \text{NM}(\theta_0, \sigma_\theta^2)$ is then given by

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

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

$$\hat{\theta} \pm \sigma \times \sqrt{\left[2\text{W}_0\left\{\frac{(1 + \sigma_\theta^2/\sigma^2)^{3/2}\sqrt{e}}{2k}\right\} - 1\right]\left(1 + \frac{\sigma^2}{\sigma_\theta^2}\right)} \tag{9}$$

with $W_0(\cdot)$ denoting the principal branch of the Lambert W function. The Lambert W function is the (complex) multivalued function $W(\cdot)$ satisfying $W(x)\exp\{W(x)\} = x$. For real x, it is defined for $x \in [-1/e, \infty)$. For $x \geq 0$ the function has a unique value, whereas in the interval $x \in (-1/e, 0)$, the function has two branches: $W_0(x) > -1$ for all $x \in (-1/e, 0)$ termed the principal branch, and $W_{-1}(x) < -1$ for all $x \in (-1/e, 0)$, see Corless et al. (1996) for more details. It is possible that the support interval (9) is empty, as for the other two types of support intervals. This happens when the Lambert W term is smaller than one half so that the square root is undefined. Since $W_0(0.82) \approx 1/2$, this situations occurs when $(1 + \sigma_\theta^2/\sigma^2)^{3/2} < 0.82 \times 2k\sqrt{e}$, meaning that the standard error σ has to be sufficiently small relative to the prior spread parameter σ_θ and the support level k, so that the interval is non-empty.

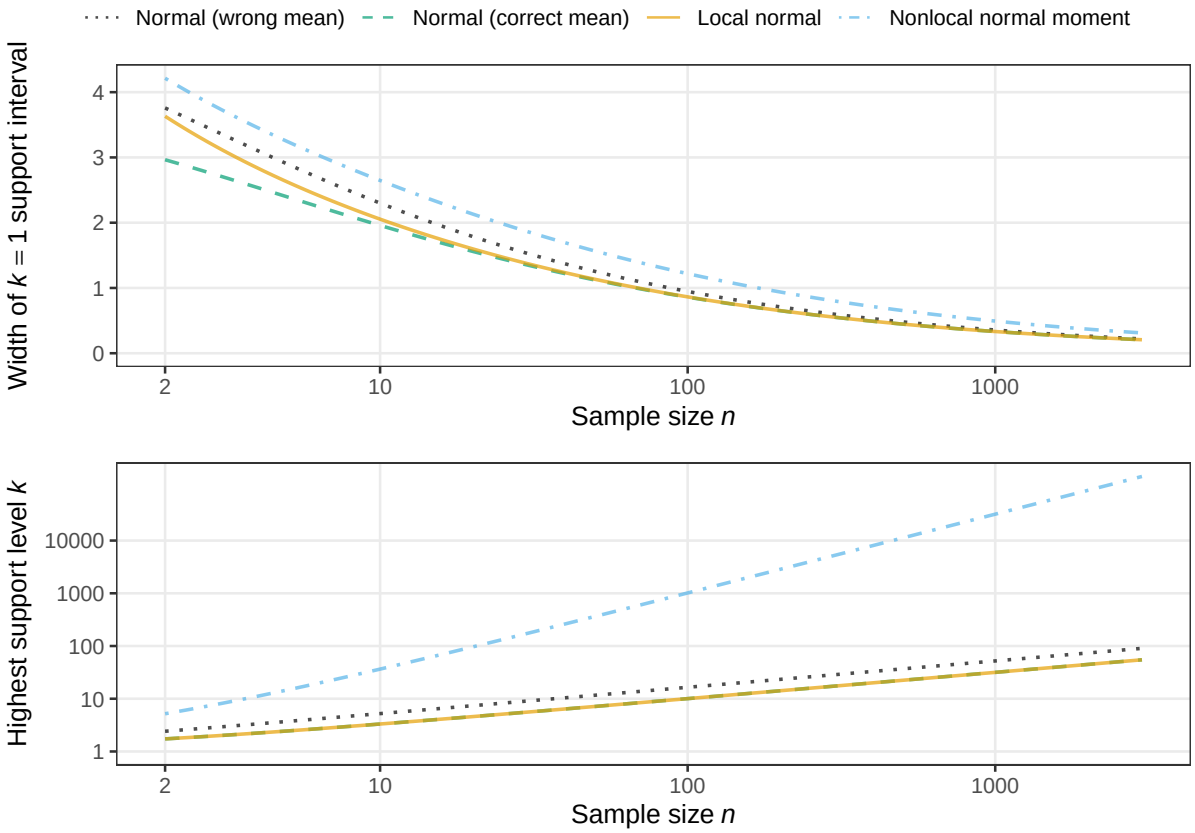


Figure 2: Comparison of prior distributions for the parameter θ under the alternative $\mathcal{H}_1$ in terms of resulting support interval width and highest level for which it is non-empty. A data model $\hat{\theta} \mid \theta \sim N(\theta, \lambda^2/n = 4/n)$ is assumed in all cases. The prior scale/spread parameter is set to $\sigma_\theta = 2$. The normal prior (correct mean) has a mean equal to the parameter estimate $\hat{\theta}$, while the normal prior (wrong mean) has a mean one standard deviation $\lambda = 2$ away from $\hat{\theta}$.

2.4 Comparison of priors

To better understand the advantages and disadvantages of the previously discussed priors, the resulting support intervals can be compared in terms of their width as a function of the sample size n (Figure 2 top). For small sample sizes, the normal prior with mean equal to the observed

parameter estimate produces the narrowest $k = 1$ support intervals, followed by the local normal prior, the normal prior with mean one standard deviation away from the observed estimate, and lastly the nonlocal normal moment prior. Thus, a well-chosen normal prior can increase the precision of support inference, whereas a poorly chosen normal prior can decrease precision. However, the differences in width between the priors mostly disappear with increasing sample size. In the realistic range between ten and a few hundred samples, the local normal prior seems to be a reasonable default choice, as it leads to support intervals almost as narrow as the normal (correct mean) prior, without the need to specify a mean.

Another aspect in which the priors can be compared is the highest support level k for which the resulting support intervals are non-empty (Figure 2 bottom). We see that for the same sample size n, the highest support levels from the normal and local normal priors are similar and show the same growth rates, while the highest support level from the nonlocal moment prior is higher and grows much faster. This is expected because nonlocal priors are designed to produce Bayes factors with faster accumulation of evidence for the null hypothesis. Thus, although nonlocal moment priors result in wider support intervals than the other priors, for small sample sizes they may be the only type of prior that can produce a support interval at, say, Jeffreys's strong evidence level $k = 10$.

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 a parameter estimate 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 strongest 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). Here we refer to these Bayes factor bounds as *minimum Bayes factors* for the null $\mathcal{H}_0$ over the alternative $\mathcal{H}_1$, as we are interested in the support for null values θ_0.

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

$$\text{minSI}_k = \{\theta_0 : \text{minBF}_{01}(x; \theta_0) \geq k\}\,, \tag{10}$$

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, given the data, 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

$$\text{minBF}_{01}(\hat{\theta}; \theta_0) = \exp\left\{-\frac{1}{2}\frac{(\hat{\theta} - \theta_0)^2}{\sigma^2}\right\}, \tag{11}$$

for which twice the negative log equals the standard likelihood ratio test statistic when $\hat{\theta}$ is the MLE. Inverting (11) for θ_0 leads to the k minimum support interval

$$\hat{\theta} \pm \sigma \times \sqrt{-2 \log k}. \tag{12}$$

Interestingly, defining a support interval relative to the likelihood of the data under the MLE has already been suggested by Fisher (1956). Table 1 shows Fisher's classification of evidence for this type of interval. Also Royall made use of the minimum support interval (12), usually with support levels $k = 1/8$ and $k = 1/32$. He noted: "The 1/8 and 1/32 likelihood intervals are not confidence intervals, in general, but they truly represent what confidence intervals are often mistaken to represent, namely parameter values that the sample does not represent evidence against, that is, values that are 'consistent with the observations'. We can speak in this way, asserting that there is not strong evidence against a point inside the interval, without reference to an alternative value, because the statement is true for all alternatives. Every point inside the 1/8 interval is consistent with the observations in the strong sense that there is no other possible value of the parameter that is better supported by a factor as large as 8" (Royall, 1997, p. 101). While we agree that the support interval (12) is a useful bound, it is important to note that from a Bayesian perspective it represents the most blatantly biased assessment of support in the sense that assigning a point prior at the observed parameter estimate hardly reflects prior knowledge about θ but can rather be considered cheating (Berger and Sellke, 1987). This is reflected by the fact that for a given estimate (i.e., data set) and fixed support level k, the interval represents the narrowest support interval among all possible support intervals. When minimizing over the class of all two-parameter normal priors, i.e., the Bayes factor (4), we also obtain the same minimum Bayes factor (11) and consequently the same minimum support interval (12).

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 $\sigma_\theta^2 = \max\{(\hat{\theta} - \theta_0)^2 - \sigma^2, 0\}$ maximizes the marginal likelihood of the data under $\mathcal{H}_1$. Plugging this variance in the Bayes factor (6) leads to the minimum Bayes factor over the class of local normal priors

$$\text{minBF}_{01}(\hat{\theta}; \theta_0) = \begin{cases} \frac{|\hat{\theta} - \theta_0|}{\sigma} \exp\left\{-\frac{(\hat{\theta} - \theta_0)^2}{2\sigma^2}\right\} \sqrt{e} & \text{if } \frac{|\hat{\theta} - \theta_0|}{\sigma} > 1 \\ 1 & \text{else} \end{cases} \tag{13}$$

as first shown by Edwards et al. (1963). Equating (13) to k and solving for θ_0 leads then to the k minimum support interval

$$\hat{\theta} \pm \sigma \times \sqrt{-\text{W}_{-1}(-k^2/e)}, \tag{14}$$

with $\text{W}_{-1}(\cdot)$ the branch of the Lambert W function that satisfies $\text{W}(y) < -1$ for $y \in (-e^{-1}, 0)$. For $k = 1$, the standard error multiplier becomes $\text{M} = \sqrt{-\text{W}_{-1}(-1/e)} = 1$. Hence, the data provide support for all parameter values within one standard error around the observed parameter estimate $\hat{\theta}$ when the class of priors for the parameter is given by local normal alternatives.

3.3 Class of *p*-based alternatives

Vovk (1993) and 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 $\text{Beta}(\xi, 1)$ 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

$$\text{minBF}_{01}(p; \theta_0) = \begin{cases} -ep \log p & \text{if } p \leq e^{-1} \\ 1 & \text{else} \end{cases} \tag{15}$$

with $p = 2\{1 - \Phi(|\hat{\theta} - \theta_0|/\sigma)\}$. Equating (15) to k and solving for θ_0, leads to the k minimum support interval

$$\hat{\theta} \pm \sigma \times \Phi^{-1}\left[1 - \frac{\exp\{\text{W}_{-1}(-k/e)\}}{2}\right]. \tag{16}$$

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

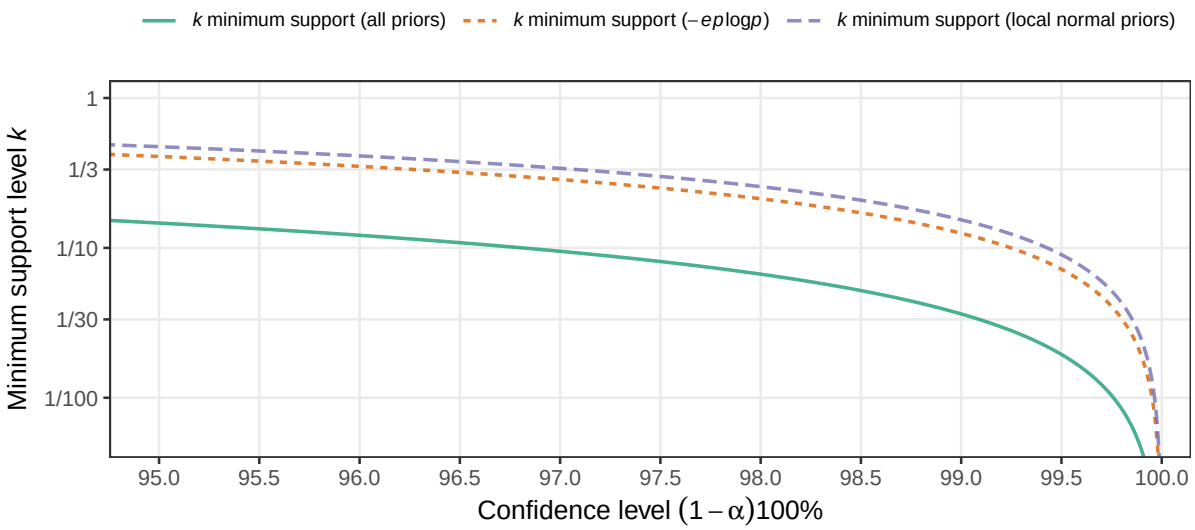


Figure 3: Mapping between confidence level $(1-\alpha)100\%$ and minimum support level k for different types of minimum support intervals.

3.4 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 minimum support level k and the confidence level $(1-\alpha)100\%$ of the approximate 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, a $k = 1/2.5$ support level for the $-ep \log p$, and a $k = 1/2.1$ support level for the local normal prior calibration. Conversely, the $k = 1/10$ minimum support interval corresponds to the 96.81% confidence interval for the class of all priors, the 99.25% confidence interval for $-ep \log p$, and the 99.43% confidence intervals for the local normal prior calibration. Similar to the mappings between Bayes factor bounds and p-values (Held and Ott, 2018), the mappings displayed in Figure 3 provide confidence intervals an evidential interpretation. Specifically, they enhance their long-term frequency interpretation with an interpretation that directly relates to the minimum support that the observed data provide for the parameter values in the interval.

4 Example RECOVERY trial

We now compute the above (minimum) support intervals for the data from the RECOVERY trial (RECOVERY Collaborative Group, 2021). With the standard error σ known, the minimum support intervals are fully specified and can be readily computed. For the normal, local normal, and the nonlocal normal moment prior we choose their parameters as follows. 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 $\mu_\theta = -0.22$ along with the unit-information variance $\sigma^2_\theta = 4$ for a log hazard ratio (Spiegelhalter et al., 2004, Section 2.4.2). Likewise, we use the

unit-information variance $\sigma_\theta^2 = 4$ as the variance of the local normal prior. The spread parameter of the nonlocal moment prior σ_θ is elicited with a similar approach as in Pramanik and Johnson (2022); The value $\sigma_\theta = 0.28$ is selected so that 90% probability mass is assigned to log hazard ratios between $\theta_0 - \log 2$ and $\theta_0 + \log 2$, representing effect sizes that at most half or double the mortality hazards relative to the null value θ_0.

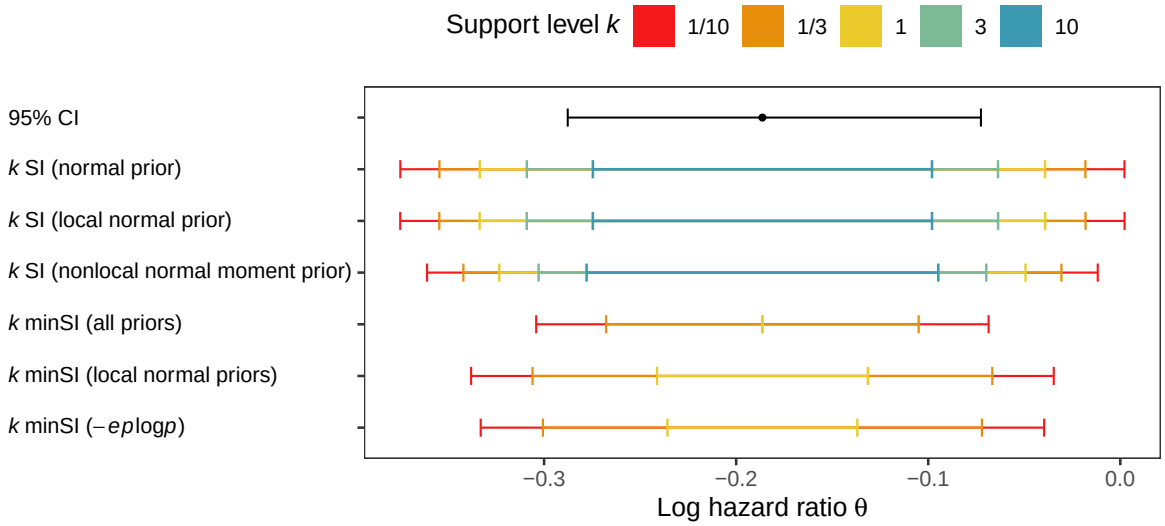


Figure 4: Different support intervals for the data from the RECOVERY trial. The normal prior is centered around $\mu_\theta = -0.22$ and has unit variance $\sigma_\theta^2 = 4$. The local normal prior also has unit variance $\sigma_\theta^2 = 4$. The spread parameter of the nonlocal normal moment prior is $\sigma_\theta = 0.28$.

Figure 4 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, each parameter value not included in a k support interval corresponds to a point-null hypothesis for which the respective Bayes factor is smaller than k, similar 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.

The three bottom rows in Figure 4 show different types of k minimum support intervals computed for the data from the RECOVERY trial. Since minimum support intervals are only non-empty 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 local normal

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 narrower than the ones based on local normal and $-ep \log p$, yet all of them are narrower than the ordinary support intervals. This illustrates that minimum support intervals provide an overly pessimistic assessment of support for parameter values, in the same way that Bayes factor bounds provide an overly pessimistic quantification of evidence for the null hypothesis.

5 Design of new studies based on support

The sample size of a future study is typically derived to achieve (i) a targeted power of a hypothesis test, or (ii) a targeted precision of a future confidence/credible interval. Here, we provide an alternative where the sample size of a future study is determined to achieve a desired level of support.

Assume we wish to conduct a study and analyze the resulting parameter estimate $\hat{\theta}$ using the support interval based on a normal prior (5). Further assume that we either specify a reasonable prior from existing knowledge or use 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 be non-empty, the standard error σ of the parameter estimate $\hat{\theta}$ needs to be sufficiently small so that the term in the square root becomes non-negative, i. e., it must hold that

$$\log\left(1 + \frac{\sigma_\theta^2}{\sigma^2}\right) + \frac{(\hat{\theta} - \mu_\theta)^2}{\sigma^2 + \sigma_\theta^2} \geq 2 \log k. \tag{17}$$

The sample size n can now be determined such that the standard error σ is small enough for (17) 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 ($\mu_\theta = \hat{\theta}$ and $\sigma_\theta^2 = n\sigma^2$), the simple expression $n \geq k^2 - 1$ mentioned earlier exists. For instance, if we want a $k = 10$ support interval to be non-empty, we must take at least $10^2 - 1 = 99$ samples.

While the previously described approach guarantees that a $k > 1$ support interval is non-empty and includes at least one parameter value θ, one may want to guarantee that the resulting k support interval will span a desired length

$$\ell = 2\sigma \times \mathrm{M}_k, \tag{18}$$

with M_k the standard error multiplier of a k support interval. In general, numerical methods are required for computing the n such that (18) is satisfied, yet again for the support interval based

on Jeffrey's approximate Bayes factor there are explicit solutions available

$$n = k^2 \exp\left\{-\mathrm{W}\left(-\frac{k^2\ell^2}{4\lambda^2}\right)\right\} \tag{19}$$

with λ^2 the variance of one (effective) observation and assuming $\log(1+n)/\log(n) \approx 1$. From (19) 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 $\ell \leq (4\lambda^2)/k^2$, (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 standard error of $\sigma^2 = \lambda^2/n$ with $\lambda^2 = 4$, a support level $k = 10$, and a desired width $\ell = 0.2$, equation (19) leads to the sample sizes $n_1 = 143$ and $n_2 = 862$ (when rounded to the next larger integer). Both lead to the $k = 10$ support interval spanning the desired width $\ell = 0.2$, yet for the study employing the larger sample size n_2 other support intervals with higher support levels k can be computed compared to a study employing the smaller sample size n_1.

6 Error control via the universal bound

The universal bound (Royall, 1997, Section 1.4) ensures that for $k < 1$ and when the null hypothesis $\mathcal{H}_0\colon \theta = \theta_0$ is true, the probability for finding evidence at most of level k for $\mathcal{H}_0$ cannot be larger than k, that is

$$\Pr\left\{\mathrm{BF}_{01}(x; \theta_0) \leq k \,|\, \mathcal{H}_0\right\} \leq k \tag{20}$$

for any prior of θ under the alternative $\mathcal{H}_1$. Remarkably, the universal bound is also valid under sequential analyses with optional stopping as soon as a Bayes factor smaller than k is obtained (Robbins (1970); Pace and Salvan (2020)). 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.

Lindon and Malek (2020) proved that k support sets with $k < 1$ are also valid $(1-k)100\%$ confidence sets. Their proof and the related "safe and anytime valid inference" theory (see e. g., Grünwald et al., 2019) is based on relatively technical results from martingale theory. We now briefly show how the universal bound can also be used to derive error rate guarantees for support intervals. Assume there is a true parameter $\theta = \theta_*$. For any (data-independent) prior for θ under the alternative hypothesis $\mathcal{H}_1$, the coverage of the corresponding k support set SI_k with $k < 1$ is bounded by

$$\begin{aligned}
\Pr\left(\mathrm{SI}_k \ni \theta_* \,|\, \theta = \theta_*\right) &= \Pr\left\{\mathrm{BF}_{01}(x; \theta_*) \geq k \,|\, \theta = \theta_*\right\} \\
&= 1 - \Pr\left\{\mathrm{BF}_{01}(x; \theta_*) < k \,|\, \theta = \theta_*\right\} \\
&\geq 1 - k
\end{aligned} \tag{21}$$

where the first equality follows from the definition of a k support set (2), whereas the inequality follows from the universal bound (20). This shows that a k support set with $k < 1$ is also a

valid $(1 - k)100\%$ confidence sets, even under sequential analyses with optional stopping, so that computing support intervals based on accumulating data leads to a $(1 - k)100\%$ confidence sequence (Lai, 1976; Howard et al., 2021). Of course, the coverage bound rests on the assumption that the data model is correctly specified and a misspecified data model will result in incorrect coverage. Furthermore, the bound is based on simple null hypotheses, but it can also be shown to hold for composite null hypotheses when special types of priors are assigned to the nuisance parameters (Hendriksen et al., 2021).

For the case of a univariate parameter θ as considered earlier, construction of $(1 - k)100\%$ approximate confidence interval via the normal prior support interval from (5) corresponds to the proposal by Pace and Salvan (2020). 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 is also applicable to parameter 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 (2020) for further details.

A $k < 1$ support interval will usually be wider than a standard $(1 - k)100\%$ confidence interval. On the other hand, a $k < 1$ support interval has at least $(1 - k)100\%$ coverage, even under optional stopping (at least for point null hypotheses as is the case here), which is not satisfied by a standard $(1 - k)100\%$ confidence interval. Due to their property of valid coverage based on arbitrary number of looks at the data, $k < 1$ support interval will also typically be wider than $(1-k)100\%$ confidence intervals adjusted via group sequential or adaptive trial methodology which are more fine-tuned to specific interim analysis strategies (Wassmer and Brannath, 2016). These strategies are, however, typically more restrictive and computationally involved compared to the flexible and easily computable $k < 1$ support intervals which we present here.

It must be noted that the coverage bound (21) only holds for support intervals but not for minimum support intervals. This is because the minimum support intervals are derived based on priors that depend on the data, which violates the assumption of the universal bound. Minimum support intervals are thus only useful for giving confidence intervals an evidential interpretation, but a k minimum support interval with $k < 1$, itself does not provide $(1-k)100\%$ coverage under optional stopping.

7 Discussion

Misinterpretations and misconceptions of confidence intervals are common (Hoekstra et al., 2014; Greenland et al., 2016). We showed how confidence intervals can be reinterpreted as minimum support intervals which have an intuitive interpretation in terms of the minimum evidence that the data provide for the included parameter values. We also obtained easy-to-use formulas for different types of support intervals for an unknown parameter based on an estimate and standard error thereof. Table 2 summarizes our results, their limitation being the reliance on the normality assumption which may be inadequate for small sample sizes. More appropriate support intervals can be obtained from considering the exact likelihood of the data instead of a normal approximation, however, typically the support interval will not be available in closed-form anymore and require the raw data rather than only the point estimate and standard error.

Table 2: Summary of confidence intervals (CI), support intervals (SI), and minimum support intervals (minSI) for an unknown parameter θ based on a parameter estimate $\hat{\theta}$ 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 the boundaries of the interval, multiply by the ratio of the standard error multipliers $\text{M}_\text{B}/\text{M}_\text{A}$, and add again $\hat{\theta}$ to the boundaries of the interval. The standard error multipliers M depend on either the confidence 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 hypothesis: mean μ_θ and variance σ^2_θ for the normal prior, variance σ^2_θ for the local normal prior, and spread σ_θ for the nonlocal normal moment prior. 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, and $\text{W}_{-1}(\cdot)$ denotes the branch that satisfies $\text{W}(y) < 1$ for $y \in (-1/e, 0)$. (Minimum) support intervals are only non-empty for support levels k for which the standard error multiplier is real-valued, i. e., the term in the square root must be non-negative and/or the argument for $\text{W}_{-1}(\cdot)$ must be in $[-1/e, 0)$. All interval types can be computed with the R package `ciCalibrate` (Appendix A).

Interval type	Standard error multiplier M
$(1-\alpha)100\%$ CI	$\Phi^{-1}(1-\alpha/2)$
k SI (normal prior)	$\sqrt{\{\log(1+\sigma^2_\theta/\sigma^2) + (\hat{\theta}-\mu_\theta)^2/(\sigma^2+\sigma^2_\theta) - 2\log k\}}$
k SI (local normal prior)	$\sqrt{[\{\log(1+\sigma^2_\theta/\sigma^2) - 2\log k\}(1+\sigma^2/\sigma^2_\theta)]}$
k SI (nonlocal normal moment prior)	$\sqrt{([2\text{W}_0\{(1+\sigma^2_\theta/\sigma^2)^{3/2}/(2ke^{-1/2})\} - 1]\{1+\sigma^2/\sigma^2_\theta\})}$
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 that 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.

It is also not clear which support level k should be used for computing support intervals. If space permits, we recommend to visualize the Bayes factor as a function of the null value as in Figure 1. A similar approach has also been proposed by Grünwald (2023) under the name of E-posterior. The Bayes factor visualization provides readers with a more gradual assessment of support, and any desired k support interval can be read off from it. If there are space constraints, a compromise is to report support intervals for different levels (e. g., $k \in \{1/10, 1, 10\}$) or to present a forest plot with "telescope" style support intervals with ascending support levels stacked on top of each other, as in Figure 4. We are hesitant to recommend a "default" support level because any classification of support is arbitrary, just like the 95% confidence level convention. We believe that $k = 1$ is perhaps the least arbitrary default level, as it represents the tipping point at which the included parameter values begin to receive support from the data (although not necessarily strong support).

Other approaches for reinterpreting confidence intervals have been proposed. 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". This notion of compatibility is logically weaker than the notion of support considered in this paper as a failure to refute a parameter value cannot establish that this parameter value is supported without reference to alternatives (Greenland, 2023). Compatibility intervals are in this sense similar to minimum support intervals; without a specified prior under the alternative hypothesis only the maximum surprisal/evidence *against* the included parameter values can be quantified.

We also showed how the coverage of k support intervals with $k < 1$ is bounded by $(1-k)100\%$, which holds even under sequential analyses with optional stopping. For instance, a $k = 1/20$ support interval has valid 95% coverage. Of course, such error rate guarantees rest on the assumption that the 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; Fong and Holmes, 2020). Such "descriptive inferential statistics" are especially important for the analysis of convenience data samples which typically violate assumptions of the underlying statistical model (Amrhein et al., 2019; Shafer, 2021). In fact, even one of the best known proponents of p-values – R.A. Fisher – noted "For all purposes, and more particularly for the *communication* of the relevant evidence supplied by a body of data, the values of the Mathematical Likelihood are better fitted to analyse, summarize, and communicate statistical evidence of types too weak to supply true probability statements" (Fisher, 1956, p. 70) clearly recognizing the importance of inferential tools based on relative likelihood for making sense out of data.

Software and data

The point estimate and 95% confidence interval of the adjusted log hazard ratio were extracted from the abstract of RECOVERY Collaborative Group (2021). All analyses were conducted in the R programming language version 4.3.0. Code and data for reproducing the results in this manuscript are available at `https://github.com/SamCH93/ECoCI`. A snapshot of the GitHub repository at the time of writing this article is archived at `https://doi.org/10.5281/zenodo.6723249`. An R package for calibration of confidence intervals to (minimum) support intervals is available at `https://CRAN.R-project.org/package=ciCalibrate`, see Appendix A for an illustration.

Acknowledgments

We thank Leonhard Held for helpful comments on an earlier version of the manuscript. We thank Michael Lindon for interesting discussions and for letting us know about his work on the

connection between support and confidence sets. We thank Glenn Shafer for attending us about R.A. Fisher's work on relative likelihood. We thank Sander Greenland for valuable feedback on the first version of the manuscript. We thank the editor Joshua Tebbs, the anonymous associate editor and the anonymous reviewer for useful comments. Our acknowledgment of these individuals does not imply their endorsement of this article. 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.

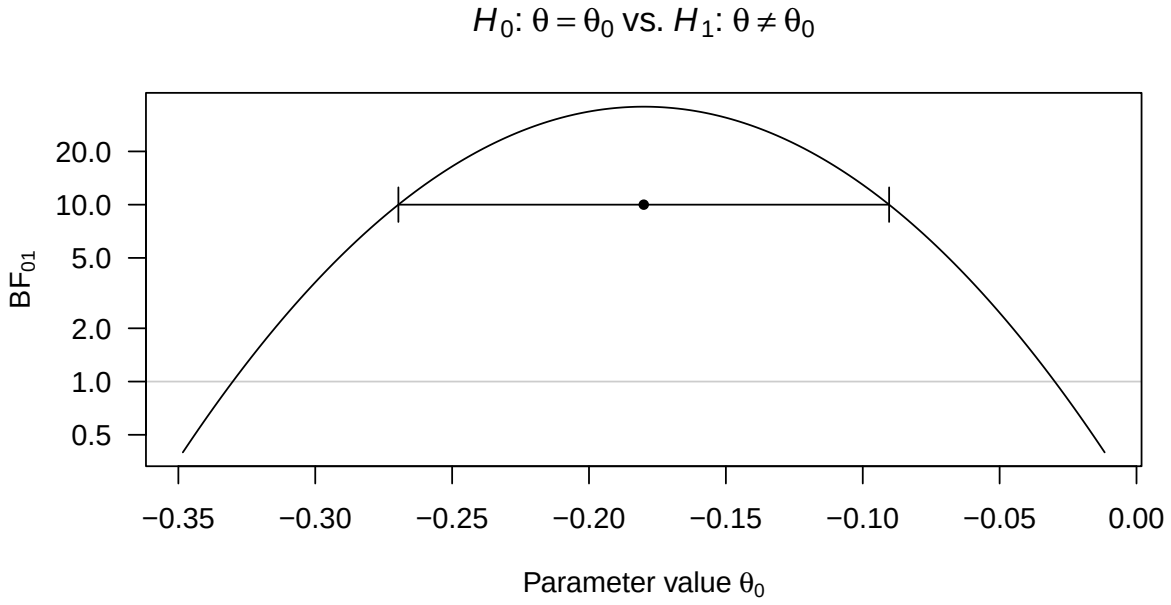
Appendix A The ciCalibrate package

We provide an R implementation of the support intervals and underlying Bayes factor functions from Table 2. The package is available at `https://CRAN.R-project.org/package=ciCalibrate` and can be installed by executing `install.packages("ciCalibrate")` in an R console. The following code snippet illustrates the computation and plotting of support interval and Bayes factor function.

```
## 95% CI from RECOVERY trial
logHRci <- c(-0.29, -0.07)
## compute a support interval with level k = 10
library("ciCalibrate") # install with install.packages("ciCalibrate")
si10 <- ciCalibrate(ci = logHRci, ciLevel = 0.95, siLevel = 10,
                    method = "SI-normal", priorMean = 0, priorSD = 2)
si10

##
## Point Estimate [95% Confidence Interval]
## -0.18 [-0.29,-0.07]
##
## Calibration Method
## Normal prior for parameter under alternative
## with mean m = 0 and standard deviation sd = 2
##
## k = 10 Support Interval
## [-0.27,-0.09]

## plot Bayes factor function with support interval
plot(si10)
```



References

Amrhein, V., Trafimow, D., and Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1):262–270. doi:10.1080/00031305.2018.1543137.

Berger, J., Bayarri, M. J., and Pericchi, L. R. (2013). The effective sample size. *Econometric Reviews*, 33(1-4):197–217. doi:10.1080/07474938.2013.807157.

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

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.

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

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.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Oliver & Boyd, Edinburgh.

Fong, E. and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496. doi:10.1093/biomet/asz077.

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(477):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. (2023). Divergence vs. 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. (2019). Safe testing. doi:10.48550/ARXIV.1906.07801. Preprint.

Grünwald, P. (2023). The E-posterior. *Philosophical Transactions of the Royal Society A*. doi:10.1098/rsta.2022.146. URL `https://arxiv.org/abs/2301.01335`.

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.

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.

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

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080. doi:10.1214/20-aos1991.

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.

Lai, T. L. (1976). On confidence sequences. *The Annals of Statistics*, 4(2). doi:10.1214/aos/1176343406.

Lindon, M. and Malek, A. (2020). Sequential testing of multinomial hypotheses with applications to detecting implementation errors and missing data in randomized experiments. URL `https://arxiv.org/abs/2011.03567v1`.

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., and Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80:40–55. doi:10.1016/j.jmp.2017.05.006.

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. (2020). 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. *BMC Medical Research Methodology*, 20(1):244. doi:10.1186/s12874-020-01105-9.

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, 27(3):411–427. doi:10.1177/0049124199027003005.

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.

Shafer, G. (2021). Descriptive probability. Working paper #59 (version September 30, 2021). `http://probabilityandfinance.com/articles/59.pdf`.

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

Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):317–341. doi:10.1111/j.2517-6161.1993.tb01904.x.

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. (2022). The support interval. *Erkenntnis*, 87(2):589–601. doi:10.1007/s10670-019-00209-z.

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.

Wassmer, G. and Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer International Publishing, Cham, Switzerland. doi:10.1007/978-3-319-32562-0.

Computational details

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

## 2023-05-10 17:02:29.194182 Europe/Zurich

sessionInfo()

## R version 4.3.0 (2023-04-21)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.6 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
##
## time zone: Europe/Zurich
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.4.2      colorspace_2.1-0   lamW_2.1.2         ciCalibrate_0.42.2
## [5] knitr_1.42
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.3       highr_0.10         dplyr_1.1.2        compiler_4.3.0
##  [5] ggsignif_0.6.4     tidyselect_1.2.0   Rcpp_1.0.10        gridExtra_2.3
##  [9] tidyr_1.3.0        scales_1.2.1       R6_2.5.1           ggpubr_0.6.0
## [13] labeling_0.4.2     generics_0.1.3     backports_1.4.1    tibble_3.2.1
## [17] car_3.1-2          munsell_0.5.0      RColorBrewer_1.1-3 pillar_1.9.0
## [21] rlang_1.1.0        utf8_1.2.3         broom_1.0.4        xfun_0.39
## [25] RcppParallel_5.1.7 cli_3.6.1          withr_2.5.0        magrittr_2.0.3
```

```
## [29] grid_4.3.0          cowplot_1.1.1       lifecycle_1.0.3     vctrs_0.6.2
## [33] rstatix_0.7.2       evaluate_0.20       glue_1.6.2          farver_2.1.1
## [37] abind_1.4-5         carData_3.0-5       fansi_1.0.4         purrr_1.0.1
## [41] tools_4.3.0         pkgconfig_2.0.3
```