The Support Interval

Wagenmakers, E.-J.; Gronau, Q.F.; Dablander, F.; Etz, A.

DOI
10.31234/osf.io/zwnxb
10.1007/s10670-019-00209-z

Publication date
2022

Document Version
Submitted manuscript

Published in
Erkenntnis

License
CC BY

The Support Interval

Eric–Jan Wagenmakers[1], Quentin F. Gronau[1], Fabian Dablander[1], and Alexander Etz[2]
1 University of Amsterdam
2 University of California, Irvine

Correspondence concerning this manuscript should be addressed to:
E.-J. Wagenmakers
University of Amsterdam, Department of Psychology
Nieuwe Achtergracht 129B
1018VZ Amsterdam, The Netherlands
E–mail may be sent to EJ.Wagenmakers@gmail.com.

Abstract

A frequentist confidence interval can be constructed by inverting a hypothesis test, such that the interval contains only parameter values that would not have been rejected by the test. We show how a similar definition can be employed to construct a Bayesian support interval. Consistent with Carnap's theory of corroboration, the support interval contains only parameter values that receive at least some minimum amount of support from the data. The support interval is not subject to Lindley's paradox and provides an evidence-based perspective on inference that differs from the belief-based perspective that forms the basis of the standard Bayesian credible interval.

In frequentist statistics, there is an intimate connection between the p-value null-hypothesis significance test and the confidence interval of the test-relevant parameter. Specifically, a $[100 \times (1-\alpha)]\%$ confidence interval contains only those parameter values that would not be rejected if they were subjected to a null-hypothesis test with level α. That is, frequentists confidence intervals can often be constructed by inverting a null-hypothesis significance test (e.g., Natrella, 1960; Stuart, Ord, & Arnold, 1999, p. 175). Thus, the construction of the confidence interval involves, at a conceptual level, the computation of p-values.

In Bayesian statistics, in contrast, there exists a conceptual divide between the Bayes factor hypothesis test and the credible interval. On the one hand, the Bayes factor (e.g., Etz & Wagenmakers, 2017; Haldane, 1932; Jeffreys, 1939; Kass & Raftery, 1995; Wrinch & Jeffreys, 1921) reflects the relative predictive adequacy of two competing models or hypotheses, say $\mathcal{H}_0$ (which postulates the absence of the test-relevant parameter) and $\mathcal{H}_1$ (which postulates the presence of the test-relevant parameter). On the other hand, under the assumption that $\mathcal{H}_1$ is true, the associated credible interval for the test-relevant

parameter provides a range that contains 95% of the posterior mass. In other words, the Bayes factor test seeks to quantify the evidence for the presence or absence of an effect, whereas the credible interval quantifies the size of the effect under the assumption that it is present. For this reason, one may encounter paradoxical situations in which the following are simultaneously true: (1) the Bayes factor supports the point hypothesis $\mathcal{H}_0 : \theta = \theta_0$ over the composite hypothesis $\mathcal{H}_1$ in which θ is assigned some continuous prior distribution; and (2) the central 95% credible interval excludes the value θ_0.

As a concrete example, consider a binomial test with $\mathcal{H}_0 : \theta_0 = 1/2$ and $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 1)$, and assume we observe 60 successes and 40 failures. Figure 1 shows that the Bayes factor slightly favors $\mathcal{H}_0 : \theta_0 = 1/2$, whereas the 95% credible interval just excludes that point.

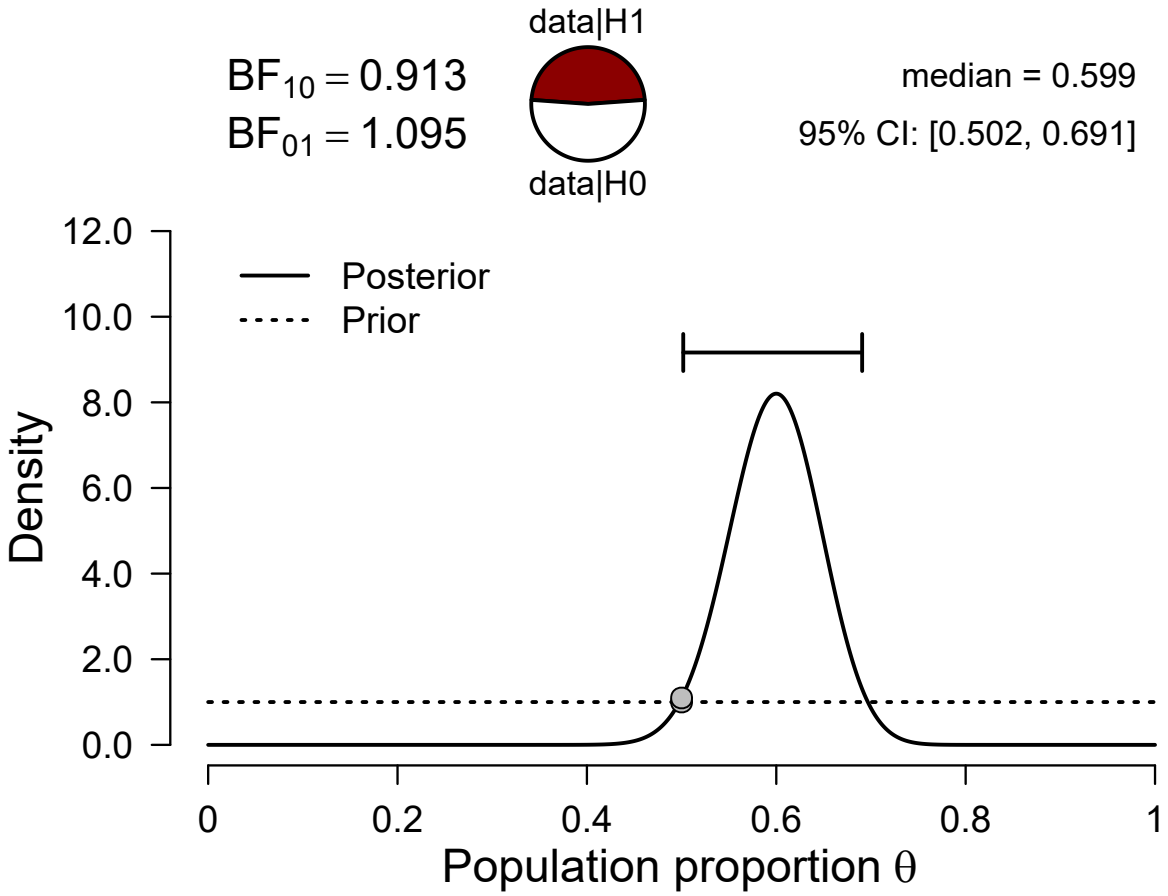

Figure 1. Based on 60 successes and 40 failures, a binomial test with $\mathcal{H}_0 : \theta_0 = 1/2$ versus $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 1)$ yields (very slight) evidence in favor of $\mathcal{H}_0 : \theta_0 = 1/2$, whereas the central 95% credible interval under $\mathcal{H}_1$ ranges from 0.502 to 0.691, just excluding the point $\theta_0 = 1/2$. Figure from JASP, `jasp-stats.org`.

There are different responses to this paradoxical state of affairs:

1. One may blame the Bayes factor, or, more specifically, one may blame the fact that the prior distribution for θ under $\mathcal{H}_1$ is overly wide, which harms predictive performance of

$\mathcal{H}_1$. However, the conflict arises irrespective of the prior distribution; that is, if person X specifies, for instance, a Beta(a, b) prior, then person Y can present a fictitious data set for which the paradox emerges. This implies that the prior for θ cannot be the cause of the conflict.

2. One may recognize that Figure 1 does not in fact present the complete posterior distribution for θ. Instead, the complete (marginal) distribution for θ consists of a posterior spike at $\theta_0 = 1/2$ under $\mathcal{H}_0$ and the continuous distribution for θ under $\mathcal{H}_1$ (e.g., Rouder, Haaf, & Vandekerckhove, 2018). Ignoring the posterior spike at θ_0 paints an overly optimistic picture of what values θ is likely to have.

3. One may realize that the paradox is a contradiction that is only apparent; thus, one may simply accept that intervals computed under $\mathcal{H}_1$ may exclude values that, when considered in isolation, remain relatively plausible.

Here we explore a fourth response, one that attempts to define a Bayesian interval based on the same principles that underlie the construction of the frequentist confidence interval. This rather unconventional interval defines a set of values of θ that predicted the observed data relatively well, and it prevents the paradoxical situation outlined above from arising. Before introducing the interval, which is based on earlier work by Keynes (1921), Carnap (1950), Evans (1997, 2015), Morey, Romeijn, and Rouder (2016), and Rouder and Morey (in press), we provide some background information on the Bayes factor.

Background on the Bayes Factor

The Bayes factor quantifies the degree to which data y change the relative prior plausibility of two hypotheses (say $\mathcal{H}_0$ and $\mathcal{H}_1$) to the relative posterior plausibility, as follows:

$$\underbrace{\frac{p(\mathcal{H}_0 \mid y)}{p(\mathcal{H}_1 \mid y)}}_{\substack{\text{Relative posterior}\\ \text{uncertainty}}} = \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\substack{\text{Relative prior}\\ \text{uncertainty}}} \times \underbrace{\frac{p(y \mid \mathcal{H}_0)}{p(y \mid \mathcal{H}_1)}}_{\substack{\text{Bayes factor}\\ \text{BF}_{01}}}. \tag{1}$$

For concreteness, consider a binomial test between $\mathcal{H}_0 : \theta_0 = 1/2$ vs. $\mathcal{H}_1 : \theta \sim \text{Beta}(2, 2)$. Suppose the data at hand consist of 8 successes and 2 failures. In this specific case, the Bayes factor is given by

$$\text{BF}_{01} = \frac{p(y \mid \theta_0 = 1/2)}{p(y \mid \theta \sim \text{Beta}(2, 2))}, \tag{2}$$

the ratio of the predictive performances for $\mathcal{H}_0$ and $\mathcal{H}_1$. But now consider only $\mathcal{H}_1 : \theta \sim \text{Beta}(2, 2)$, and observe how the data change the relative plausibilities of the different values of θ under $\mathcal{H}_1$:

$$\underbrace{p(\theta \mid y, \mathcal{H}_1)}_{\substack{\text{Posterior distribution}\\ \text{under } \mathcal{H}_1}} = \underbrace{p(\theta \mid \mathcal{H}_1)}_{\substack{\text{Prior distribution}\\ \text{under } \mathcal{H}_1}} \times \underbrace{\frac{p(y \mid \theta, \mathcal{H}_1)}{p(y \mid \mathcal{H}_1)}}_{\substack{\text{Predictive}\\ \text{updating factor}}}. \tag{3}$$

The updating factor in Equation 3, assessed for the value $\theta = 1/2$, is identical to the Bayes factor in Equation 1. In other words, when we consider only $\mathcal{H}_1$, and evaluate the change from prior to posterior ordinate at a specific θ_0, we may equally well interpret this as the Bayes factor for $\mathcal{H}_0 : \theta = \theta_0$ vs. $\mathcal{H}_1 : \theta \sim \text{Beta}(2, 2)$. This relation holds generally (e.g., Dickey & Lientz, 1970; Wetzels, Grasman, & Wagenmakers, 2010; but see Marin & Robert, 2010 and Verdinelli & Wasserman, 1995). Thus, "strength of evidence for a parameter value is precisely the relative gain in predictive accuracy when conditioning on it" (Rouder & Morey, in press). Specifically, we can rewrite Equation 3 as

$$\frac{p(\theta \mid y, \mathcal{H}_1)}{p(\theta \mid \mathcal{H}_1)} = \frac{p(y \mid \theta, \mathcal{H}_1)}{p(y \mid \mathcal{H}_1)}, \tag{4}$$

which shows that the ratio of posterior to prior density for a parameter value is precisely equal to its predictive updating factor.

To underscore this key point, Figure 2 highlights the changes from prior to posterior distribution for the above binomial example; if the posterior ordinate for a specific θ_0 is higher than the prior ordinate, the data have made that θ_0 more credible than it was before: the updating factor exceeded 1, meaning that θ_0 predicted the observed data better than average (Morey et al., 2016; Wagenmakers, Morey, & Lee, 2016).

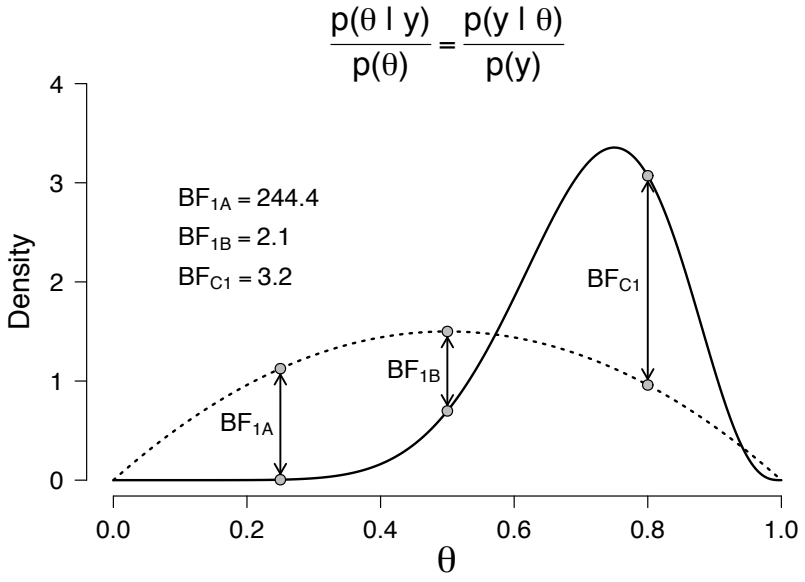

Figure 2. In Bayesian parameter estimation, the plausibility update for a specific value of θ (e.g., θ_0) is mathematically identical to a Bayes factor against a point-null hypothesis $\mathcal{H}_0 : \theta = \theta_0$. In this example, θ is assigned a Beta$(2, 2)$ prior distribution (i.e., the dotted line), the data y consist of 8 successes out of 10 trials, and the resulting posterior for θ is a Beta$(10, 4)$ distribution. Note the similarity to the Savage-Dickey density ratio test (e.g., Dickey & Lientz, 1970; Wetzels et al., 2010).

The Support Interval

As illustrated in Figure 2, some values of θ received support from the data –the updating factor was in their favor– whereas other values of θ are undermined by the data; for these values, the posterior ordinate is lower than the prior ordinate, signaling a loss in credibility. We can use this information to define an interval containing only those values of θ that receive a certain minimum level of corroboration from the data. This leads to the following definition.

Definition of the BF = k Support Interval: A BF = k support interval for a parameter θ contains only those values for θ which predict the observed data y at least k times better than average; these are values of θ that are associated with an updating factor $p(y \mid \theta)/p(y) \geq k$.

Example: Choosing a value of k

The definition of the support interval makes it apparent that in practice one must choose a value for the critical updating factor k. The choice of k depends on what we want our interval to convey about the evidence in the data. Consider again the binomial scenario illustrated in Figure 2 and suppose we seek a BF = 1 support interval for θ, that is, an interval that contains only those values whose credibility is not decreased by observing the data. This interval contains all values for θ where the posterior ordinate is equal to or exceeds the prior ordinate, and serves as a natural default choice for k. In this case, the interval ranges from $\theta \approx 0.57$ to $\theta \approx 0.94$.

We may seek an interval of values for θ that enjoy more impressive support from the data. This interval is a smaller subset of the initial BF = 1 interval. For instance, choosing $k = 3$ would produce an interval that contains all parameter values that receive at least "moderate" support from the data (according to conventions set by Jeffreys, 1939). In our binomial example, the BF = 3 support interval for θ ranges from $\theta \approx 0.75$ to $\theta \approx 0.84$. On the other hand, by choosing $k < 1$ we may also broaden our interval to encompass those values that are not strongly contraindicated by the data. The interpretation of such intervals would be analogous to how a frequentist confidence interval contains all the parameter values that would not have been rejected if tested at level α. For instance, a BF = 1/3 support interval encloses all values of θ for which the updating factor is not stronger than 3 *against*; in our example this interval ranges from $\theta \approx 0.47$ to $\theta \approx 0.97$.

Comparison to the Credible Interval

The support interval is based on evidence –how the data change our beliefs– whereas the credible interval is based on the posterior beliefs directly. Because evidence and belief are different concepts, it is straightforward to present situations in which the two intervals yield different results.

For instance, the top panel in Figure 3 shows an example of unexpected data: a Beta$(10, 3)$ prior distribution (dotted line) for a binomial parameter θ is updated to a posterior distribution (solid line) after having observed $y = 3$ successes out of $n = 20$ trials. In order to underscore that the data are unexpected under the prior, the panel also presents the likelihood (dashed line). The posterior is a compromise between prior and likelihood that is blind to any conflict between them; specifically, the exact same posterior

(and, consequently, the exact same 95% central credible interval) for θ would have resulted if θ had been assigned, say, a Beta$(5, 8)$ prior distribution and $y = 8$ successes out of $n = 20$ trials had been observed. The support interval, in contrast, is sensitive to the unexpected nature of the data. Specifically, a BF $= 1$ support interval for θ comprises all values of θ that predict the data at least as well as average: these are the values for θ where the posterior distribution equals or exceeds the prior distribution, which happens here even for a relatively wide range of values of θ.

The bottom panel in Figure 3 shows an example of relatively uninformative data: a Beta$(10, 10)$ prior distribution (dotted line) is updated to a posterior distribution (solid line) based on having observed a single success out of two trials. The 95% credible interval is relatively wide, indicating substantial uncertainty about the true value of θ; in contrast, the BF $= 1$ support interval is relatively narrow, as relatively few values of θ predicted the data better than average. For a deeper understanding of the source of the discrepancies we now take a closer look at the likelihood.

A Likelihood Perspective

The construction of the support interval is based on the change from the prior to the posterior distribution, that is,

$$\frac{p(y \mid \theta)}{p(y)} = \frac{p(y \mid \theta)}{\int p(y \mid \theta) p(\theta) \, \mathrm{d}\theta}. \tag{5}$$

The denominator is the marginal likelihood – a constant number that does not depend on θ, so that the updating factor can also be written as $c \cdot p(y \mid \theta)$. Figure 4 shows the updating factor function from the second binomial example (see Figure 2). The construction of the BF $= k$ support interval involves, first, the selection of a threshold level of evidence, say BF $= 1$ (marked in the figure with a dotted horizontal line), and then the identification of the values of θ for which the function exceeds that threshold (i.e., the values of θ in between the two gray dots that mark the intersection of the threshold with the updating factor function). Higher evidence thresholds mean smaller intervals; for instance, in Figure 4 the BF $= 3$ support interval ranges from approximately $\theta = .75$ to $\theta = .84$. If the evidence threshold is raised to 3.5 or higher, an empty interval is obtained, indicating that the data do not support any value of θ this strongly.

For a likelihoodist, the update factor function is simply a representation of the likelihood function, thus conferring the support interval the invariance properties enjoyed by likelihood-based inferences (Edwards, 1992; Royall, 1997). However, for a likelihoodist the marginal likelihood constant is arbitrary, and the update factor function may therefore be arbitrarily rescaled (e.g., to have maximum 1) without changing the inference (Etz, 2018). Consequently, a likelihood interval (e.g., Cumming, 2014; Hudson, 1971; Royall, 1997) cannot be constructed by reference to any horizontal line. Instead, an interval may be constructed by comparing the maximum height of the function to the height at any other point. For instance, a likelihood ratio interval of 3 (i.e., LR $= 3$) would contain all values of θ for which the likelihood ratio against the maximum is less than 3, that is, $p(y \mid \hat{\theta})/p(y \mid \theta) < 3$ (where $\hat{\theta}$ denotes the maximum likelihood estimate).

Consider again the case of 8 successes in 10 binomial trials. The maximum likelihood estimate is $\hat{\theta} = .8$. To obtain the likelihood ratio interval we can find the two boundary

values of θ whose likelihood ratio against $\theta = .8$ is equal to 3, which gives an interval from approximately $\theta = .58$ to $\theta = .94$. This likelihood ratio interval differs markedly from the BF = 3 interval, which ranged from $\theta = .75$ to $\theta = .84$.

Therefore, even though the LR interval and the support interval are based on the same updating/likelihood function, the intervals differ: the support interval is based on a comparison to an average, whereas the likelihood interval is based on a comparison to a maximum.

Conceptual Advantages of the Support Interval

The support interval is a unique, transformation-invariant interval that generalizes to situations with multiple parameters of interest in a straightforward fashion (Dickey & Lientz, 1970; Wetzels et al., 2010). The main advantage of the support interval, however, is conceptual: it quantifies directly which values of θ are supported by the data. Specifically, those values of θ that predict the data at least k times better than average are part of the BF = k support interval. This definition of an interval for θ prevents the interval-versus-testing paradox from arising.

The interval-versus-testing paradox that we present here can be seen as an alternative interpretation of the famous Lindley paradox (Jeffreys, 1939; Lindley, 1957). Lindley's paradox states that one can simultaneously have a frequentist test at level α reject the null hypothesis while at the same time the corresponding Bayesian test overwhelmingly supports the null hypothesis. Whereas this paradox is traditionally used to highlight the inevitable divergence of p-values and Bayesian posterior probabilities (or Bayes factors) for hypothesis testing, there is an alternative interpretation of the paradox as a warning against use of improper priors for Bayesian testing (see Bartlett, 1957, DeGroot, 1982, and Robert, 2014). However, the duality of p-values and confidence intervals suggests yet another reinterpretation of the paradox, namely, that of a divergence between confidence intervals and Bayesian hypothesis tests. In turn, because most Bayesian credible intervals are approximately confidence intervals (which converge asymptotically) Lindley's paradox can be seen highlighting the divergence of Bayesian hypothesis tests and conventional interval estimation more broadly.

Reconsider our first binomial example, shown in Figure 1, featuring 60 successes and 40 failures and a Beta$(1, 1)$ distribution for θ under $\mathcal{H}_1$. In this scenario, a BF = 1 interval ranges from $\theta \approx .498$ to $\theta \approx .697$, and a BF = 1/3 interval ranges from $\theta \approx .475$ to $\theta \approx .717$. The BF = 1 interval includes $\theta = 1/2$, indicating that the data have increased its plausibility and it should therefore not be excluded from consideration; the paradoxical difference between conclusions drawn from the interval estimate and the Bayes factor hypothesis test no longer exists.

Nuisance Parameters

The presence of nuisance parameters can pose a challenge for the Savage-Dickey representation of the Bayes factor, and thus for the interpretation of a support interval as containing those parameters which would result in a BF = k test if they were used as the test value of θ. Consider a case where there is one parameter of interest θ and a vector of nuisance parameters ϕ. Bayes' theorem dictates that after observing data y, the joint

posterior of ϕ and θ under the alternative is given by

$$p(\theta, \phi \mid y, \mathcal{H}_1) = p(\phi, \theta \mid \mathcal{H}_1) \times \frac{p(y \mid \theta, \phi, \mathcal{H}_1)}{p(y \mid \mathcal{H}_1)}, \tag{6}$$

where the final term is a joint updating factor for pairs of θ and ϕ values. It would seem that a support interval for θ (the parameter of actual interest) could be obtained by marginalizing ϕ out of both the joint posterior and joint prior, and computing the marginal updating factor for θ as in Equation 4 using the marginal posterior and prior of θ. While it is true that a value of θ contained in a k-support interval constructed in this way is indeed one which marginally becomes k-times more plausible, it will not necessarily correspond to a BF $= k$ test. Thus, in the presence of nuisance parameters a support interval does not necessarily correspond to an inversion of the Bayes factor hypothesis test.

For the equivalence of the Bayes factor in favor of θ_0 and the predictive updating factor for $\theta = \theta_0$ under the alternative to hold in the presence of nuisance parameters ϕ, we must ensure that marginalization of ϕ from both models yields $p(y \mid \theta_0, \mathcal{H}_0) = p(y \mid \theta = \theta_0, \mathcal{H}_1)$. That is, it must be true that

$$\int_\Phi p(y \mid \theta_0, \phi, \mathcal{H}_0) p(\phi \mid \mathcal{H}_0) \, \mathrm{d}\phi = \int_\Phi p(y \mid \theta = \theta_0, \phi, \mathcal{H}_1) p(\phi \mid \theta = \theta_0, \mathcal{H}_1) \, \mathrm{d}\phi. \tag{7}$$

The above equality will hold if and only if the prior distribution of ϕ under the null model matches the conditional prior distribution for ϕ given $\theta = \theta_0$ under the alternative model, that is, if and only if $p(\phi \,|\, \mathcal{H}_0) = p(\phi \,|\, \theta = \theta_0, \mathcal{H}_1)$ (Dickey, 1971). Verdinelli and Wasserman (1995) show that whenever these priors do not match, the Bayes factor and marginal updating factor will be off by a multiplicative constant, say c. Thus, if we are not careful to construct priors on nuisance parameters in just the right fashion, we will have a set of θ values in a k-support interval which correspond to BF $= c \cdot k$ test.

One way to satisfy this condition on the priors for ϕ is to take ϕ and θ as *a priori* independent under the alternative hypothesis, that is, $p(\phi, \theta \,|\, \mathcal{H}_1) = p(\phi \,|\, \mathcal{H}_1) p(\theta \,|\, \mathcal{H}_1)$. Subsequently one can directly set $p(\phi \,|\, \mathcal{H}_1)$ equal to $p(\phi \,|\, \mathcal{H}_0)$. However, in many modeling contexts the construction of independent priors can be difficult – and sometimes undesirable. For instance, Heck (2018) demonstrates that multivariate Cauchy priors do not generally satisfy the marginal-conditional condition above.

Earlier Work

The key relation between strength of evidence and relative predictive performance, expressed in Equation 4 above, was previously discussed by Carnap (1950, pp. 326-333), who called it the "general division theorem". More specifically, Carnap termed the ratio of posterior plausibility to prior plausibility the "relevance quotient", and this quotient was a critical component in Carnap's theory of confirmation: a datum D supports hypothesis $\mathcal{H}$ if and only if $P(\mathcal{H} \,|\, D) > P(\mathcal{H})$, that is, if and only if $P(\mathcal{H} \,|\, D)/P(\mathcal{H}) > 1$. Still earlier, the predictive updating factor was discussed by Keynes (1921), who called it the "coefficient of influence" (p. 170; as acknowledged by Carnap). Keynes attributed his coefficient of influence to a set of unpublished notes provided to him by W. E. Johnson, stating that his exposition relating to the coefficient of influence is "derived in its entirety from his

[Johnson's] notes" (p. 170). Carnap, Keynes, and Johnson were all considering how the data impact our belief in a singular claim or hypothesis and did not discuss the possibility of extending these ideas into an estimation context, although we should stress that this is a small step from their original ideas.

Our proposal to construct intervals based on the relative support lent by the data is similar to a more recent proposal by Evans (1997, 2015). Evans first puts forward a method for point estimation which amounts to choosing the parameter value which maximizes the posterior to prior ratio (i.e., the updating factor).[1] To quantify the uncertainty in this point estimate, Evans then constructs a "relative surprise" (or "relative belief") interval for θ that contains $\gamma\%$ of the posterior mass, such that any value in the interval has a higher updating factor than any value outside the interval (see also Shalloway, 2014).

The relative surprise interval is similar to a traditional credible interval in that it contains a fixed, predetermined proportion of the posterior mass. Thus, a 95% relative surprise interval for θ is constructed by finding a set of θ values such that (i) the posterior probability of this set is 95%, and (ii) any values not in the set have smaller updating factor than those contained in the set. The construction of a relative surprise interval is not unlike that of a $\xi\%$ highest-density interval for θ, which is constructed such that (i) the posterior probability of the interval is $\xi\%$, and (ii) any values within the interval have higher posterior density than any values outside the interval. In fact, because the updating factor is proportional to the likelihood function, the relative surprise interval for a one-to-one transformation of θ, say $\psi = g(\theta)$, is equivalent to a highest-density interval whenever the prior distribution induced for ψ is uniform.

Clearly, the relative surprise interval and our proposed support interval are closely related. Both intervals have the property that any values inside the interval have a larger updating factor than those outside the interval, and both intervals are invariant under smooth reparameterization. It is straightforward to shift the interpretation of a relative surprise interval into a support interval, and vice-versa. To determine what relative surprise coefficient γ corresponds to a BF $= k$ support interval, one can simply find the posterior probability contained in the support interval. For instance, the BF $= 3$ support interval for our second binomial example ranges from $\theta = .75$ to $\theta = .84$, and the posterior probability of that interval is .274; hence, this BF $= 3$ support interval is a $\gamma = 27.4\%$ relative surprise interval. Likewise, computing the updating factor corresponding to the boundary points of a relative surprise interval gives the critical value of a support interval.

The important difference between the two intervals is that the set of θ values in a support interval is defined by a critical value of the updating factor, so the proportion of the posterior distribution in a support interval is not fixed in advance. Indeed, a given choice of critical updating factor can even result in an empty support interval; for instance, this would occur in our second binomial example when the critical updating factor is taken to be 3.5. In contrast, a $\gamma\%$ relative surprise interval will always have posterior probability γ. For relatively large γ (e.g., $\gamma = .95$), this necessitates including values of θ in the relative surprise interval which have an updating factor smaller than one. In other words, the interval can

[1] It would appear this procedure corresponds to using the MLE as an estimate of θ because the updating factor is proportional to a likelihood function, at least in problems without nuisance parameters. However, this correspondence may not agree in general for inferences involving non-invertible functions of the model parameters (e.g., Example 1 in Evans, 1997).

include parameter values that the data have undermined. The relative surprise interval is more a summary of the posterior distribution, whereas the support interval is more a summary of the evidence in the data. This behavior of the relative surprise interval has led Evans (2015) to recommend reporting instead a so-called *plausible interval*, which only contains parameter values which have evidence in their favor – in other words, a support interval for $k = 1$. These ideas have also been expanded upon by Baskurt, Evans, et al. (2013) in the context of evidence calibration, and by Evans and Tomal (2018) in the context of multiple testing and sparsity.

More recently, Rouder and Morey (in press) have argued that the updating factor should receive more attention when teaching Bayes' rule. They display an example of an updating factor function, and mention that in their teaching, "we ask students to find intervals where the data have decreased the plausibility by more than 10-1." Similar remarks can be found in Morey et al. (2016), where their Figure 2 shows an example of a BF = 1 support interval. In sum, the support interval discussed here is a more elaborate description that is inspired by earlier work conducted by Keynes, Carnap, Evans, and Rouder, Romeijn, & Morey.[2]

Concluding Comments

The support interval is based on evaluating, for each parameter value, the degree to which it predicted the observed data better than the average prediction across all parameter values. One may argue that by omitting $p(\theta)$ and focusing entirely on the evidence that is provided by the data, the support interval is not sufficiently Bayesian; on the other hand, one may argue that the quantify $\int p(y \mid \theta)p(\theta)\,\mathrm{d}\theta$, the marginal likelihood or average predictive performance across θ, is too dependent on $p(\theta)$ – this is the common objection to Bayes factor model selection (e.g., Liu & Aitkin, 2008).

All intervals come with assumptions, limitations, and advantages, and we believe it is useful to know which parameter values have received more than a specific level of corroboration from the data. Alternatively, one could of course forgo the computation of an interval altogether and simply plot the prior and the posterior distributions (e.g., Figure 1). However, the forces of habit or nature are likely to lead researchers to extract, by eye, the intervals of interest. One interval that catches the eye is the credible interval, which is based purely on the posterior; another interval that stands out is the support interval, the collection of parameter values for which the posterior height and the prior height differ by a specific factor.

Despite its intuitive appeal, the support interval has received scant attention in the Bayesian literature on estimation. One may speculate that objective Bayesians perhaps undervalue their prior distribution, whereas subjective Bayesians overvalue it. Regardless of why measures of support in estimation have been spurned for so long, we believe that practical and theoretical considerations suggest that the support interval can provide a useful summary of what was learned from the data.

[2] After this work was completed we learned that Wiebe Pestman, Wolf Vanpaemel, and Francis Tuerlinckx had developed the same idea, but never published it (Wolf Vanpaemel, personal communication).

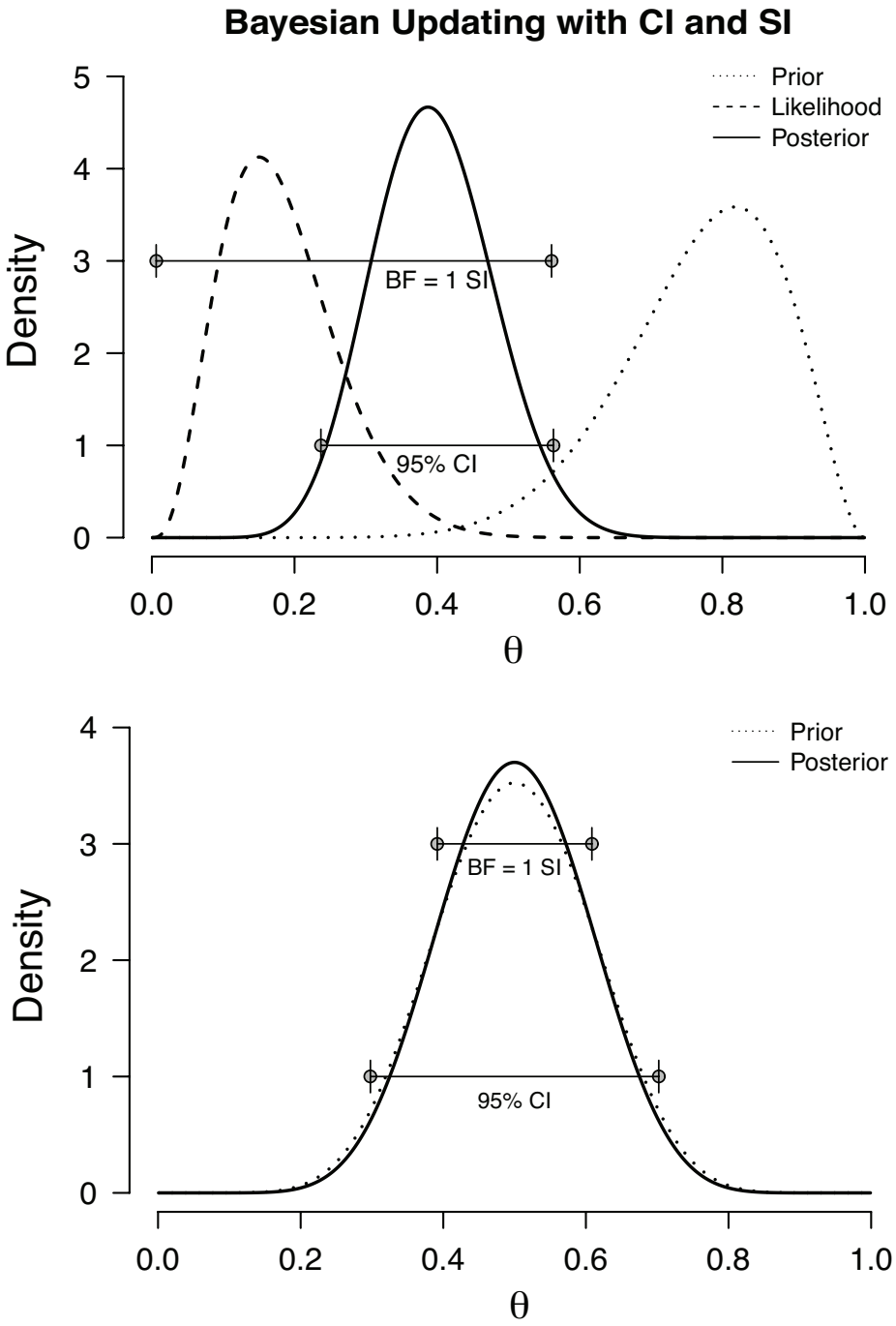

Figure 3. Differences between the support interval and the credible interval. The top panel shows a Beta$(10, 3)$ prior distribution (dotted line) which is updated to a posterior distribution based on observing $y = 3$ successes out of $n = 20$ trials; the BF $= 1$ support interval is much larger than the central 95% credible interval. The bottom panel shows a Beta$(10, 10)$ prior distribution which is updated to a posterior distribution based on observing $y = 1$ successes out of $n = 2$ trials; the BF $= 1$ support interval is smaller than the central 95% credible interval. See text for details.

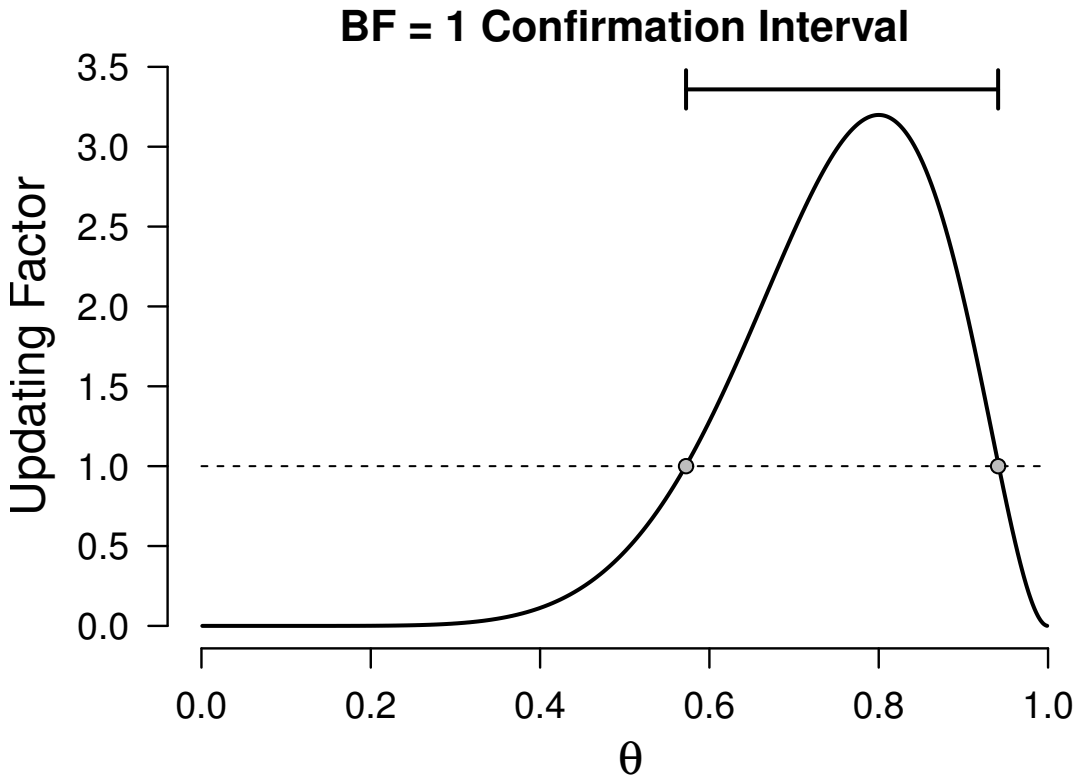

Figure 4. Example of an update factor function that quantifies the change from prior to posterior ordinate for binomial rate parameter θ. Data y consist of 8 successes and 2 failures, and θ is assigned a Beta$(2, 2)$ distribution, as in Figure 2.

References

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534.

Baskurt, Z., Evans, M., et al. (2013). Hypothesis assessment and inequalities for bayes factors and relative belief ratios. *Bayesian Analysis*, *8*(3), 569–590.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: The University of Chicago Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-?29.

DeGroot, M. H. (1982). Lindley's paradox: Comment. *Journal of the American Statistical Association*, *77*, 336–339.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*, 204–223.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Edwards, A. W. F. (1992). *Likelihood*. Baltimore, MD: The Johns Hopkins University Press.

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, *1*, 60–69.

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329.

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

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

Evans, M., & Tomal, J. (2018). Measuring statistical evidence and multiple testing. *FACETS*, *3*(1), 563–583.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *28*, 55–61.

Heck, D. (2018). A caveat on the Savage-Dickey density ratio: The case of computing Bayes factors for regression parameters. *PsyArXiv preprint*. doi: 10.31234/osf.io/7dzsj

Hudson, D. J. (1971). Interval estimation from the likelihood function. *Journal of the Royal Statistical Society. Series B (Methodological)*, *33*, 256–262.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan & Co.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Marin, J.-M., & Robert, C. P. (2010). On resolving the Savage-Dickey paradox. *Electronic Journal of Statistics*, *4*, 643–654.

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18.

Natrella, M. G. (1960). The relation between confidence intervals and tests of significance: A teaching aid. *The American Statistician*, *14*, 20–22.

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

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*, 102–113.

Rouder, J. N., & Morey, R. D. (in press). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*.

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

Shalloway, D. (2014). The evidentiary credible region. *Bayesian Analysis*, *9*, 909–922.

Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics vol. 2A: Classical inference & the linear model (6th ed.)*. London: Arnold.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, *90*, 614–618.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio test. *Computational Statistics & Data Analysis*, *54*, 2094–2102.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.