Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–16

Defining a Credible Interval Is Not Always Possible with "Point-Null" Priors: A Lesser-Known Correlate of the Jeffreys-Lindley Paradox

Harlan Campbell* and Paul Gustafson†

Abstract. In many common situations, a Bayesian credible interval will be, given the same data, very similar to a frequentist confidence interval, and researchers will interpret these intervals in a similar fashion. However, no predictable similarity exists when credible intervals are based on model-averaged posteriors whenever one of the two nested models under consideration is a so called "point-null". Not only can this model-averaged credible interval be quite different than the frequentist confidence interval, in some cases it may be undefined. This is a lesser-known correlate of the Jeffreys-Lindley paradox and is of particular interest given the popularity of the Bayes factor for testing point-null hypotheses.

1 Introduction

Recently, Bayesian tests using Bayes factors have been proposed as alternatives to frequentist hypothesis testing; see Heck et al. (2023) for a review. When using the Bayes factor (or the posterior model odds) for testing, it is often recommended that researchers also report parameter estimates and their credible intervals (e.g., Keysers et al. (2020)). Indeed, following a controversial debate about the strict binary nature of statistical tests, many now call for an additional focus on parameter estimation with appropriate uncertainty estimation; see Wasserstein and Lazar (2016).

In Campbell and Gustafson (2022), we considered how Bayesian testing and estimation can be done in a complimentary manner and concluded that if one reports a Bayes factor comparing two models, then one should also report a *model-averaged* credible interval (i.e., one based on the posterior averaged over the two models under consideration). Researchers who follow this recommendation can obtain credible intervals congruent with their Bayes factor, thereby obtaining suitable uncertainty estimation.

In many familiar situations, a posterior credible interval will be, given the same data, very similar to a frequentist confidence interval and researchers will interpret these intervals in a similar fashion; see Albers et al. (2018). However, when comparing two models, one of which involves a so-called "point-null", it is less clear whether or not such similarity can be assumed.

*Department of Statistics, University of British Columbia, harlancampbell@gmail.com
†Department of Statistics, University of British Columbia

 https://doi.org/10.1214/23-BA1397

Previous work has examined the properties of Bayesian credible intervals and how they relate to frequentist confidence intervals under various prior specifications (e.g., Casella and Berger (1987), Datta and Ghosh (1995), Greenland and Poole (2013), Held (2020)). In this paper, on the basis of a few simple examples, we will examine properties specific to model-averaged credible intervals. We will show that, when one of the two models under consideration is a point-null model, not only can the model-averaged credible interval be quite different than the confidence interval, oftentimes, for a desired probability level, it may be undefined. This is perhaps an unexpected correlate of the Jeffreys-Lindley paradox, the most well known example of the rift between frequentist and Bayesian statistical philosophies; see Wagenmakers and Ly (2021). The limitations/particularities of working with point-null models are of particular interest given the recent popularity of the Bayes factor for testing point-null hypotheses.

We begin in Section 2 by re-visiting an example of two Normal models considered previously by Wagenmakers and Ly (2021) in their discussion of the Jeffreys-Lindley paradox. In Section 3, we extend this example to consider the consequences of specifying a point-null model. We conclude in Section 4 with thoughts on the consequences, with respect to parameter estimation, of specifying point-null models.

2 A mixture of two normals

Let θ be the parameter of interest for which there are two *a priori* probable models: M_0 and M_1, defined by two different priors $\pi_0(\theta)$ and $\pi_1(\theta)$. The posterior density which appropriately acknowledges the uncertainty with regards to which of the two models is correct is the mixture density:

$$\pi(\theta|data) = \Pr(M_0|data)\pi_0(\theta|data) + \Pr(M_1|data)\pi_1(\theta|data), \tag{1}$$

where the model-specific posteriors, $\pi_0(\theta|data)$ and $\pi_1(\theta|data)$, are weighted by their posterior model probabilities, $\Pr(M_0|data)$ and $\Pr(M_1|data)$; see, for instance, Campbell and Gustafson (2022). Note that this "mixture" posterior is obtained as a result of specifying the "mixture" prior:

$$\pi(\theta) = \Pr(M_0)\pi_0(\theta) + \Pr(M_1)\pi_1(\theta), \tag{2}$$

where $\Pr(M_0)$ and $\Pr(M_1)$ are the *a priori* model probabilities.

As an example, consider two *a priori* equally probable Normal models, $M_0 : \theta \sim N(0, g_0)$ and $M_1 : \theta \sim N(0, g_1)$, such that $\Pr(M_0) = \Pr(M_1) = 0.5$. The prior density functions for the two models are defined as:

$$\pi_0(\theta) = f_{Normal}(\theta, 0, g_0), \tag{3}$$

and

$$\pi_1(\theta) = f_{Normal}(\theta, 0, g_1),$$

where $f_{Normal}(x, \mu, \sigma^2)$ is the Normal probability density function evaluated at x, with mean parameter μ and variance parameter σ^2. Let y_i be the i-th data-point, for $i =$

$1, \ldots, n$; let $\bar{y} = \sum_{i=1}^{n} y_i/n$ be the sample mean; and suppose these data are normally distributed with known unit variance such that:

$$\Pr(data|\theta) = \prod_{i=1}^{n} f_{Normal}(y_i, \theta, 1).$$

Then the Bayes factor is:

$$\text{BF}_{01} = \sqrt{\frac{1+ng_1}{1+ng_0}} \times \exp\Big(\frac{(g_0-g_1)nz^2}{2(1+ng_0)(1+ng_1)}\Big),$$

where $z = \sqrt{n}\bar{y}$. The posterior model probabilities can be calculated from the Bayes factor as:

$$\Pr(M_0|data) = \frac{\Pr(M_0)}{\Pr(M_1)/\text{BF}_{01} + \Pr(M_0)} \quad \text{and} \quad \Pr(M_1|data) = 1 - \Pr(M_0|data). \quad (4)$$

Finally, the model specific posteriors are defined as:

$$\pi_j(\theta|data) = f_{Normal}\Big(\theta, \frac{zg_j}{\sqrt{n}(\frac{1}{n}+g_j)}, \frac{g_j}{1+g_jn}\Big),$$

for $j = 0, 1$.

Having established all the components of (1), let us now consider how to define a credible interval based on the model-averaged posterior. An upper one-sided $(1-\alpha)\%$ credible interval is defined as:

$$\text{one-sided } (1-\alpha)\%\text{CrI} = [\theta^*, \infty),$$

where θ^* satisfies the following equality:

$$\Pr(\theta < \theta^*|data) = \alpha. \quad (5)$$

Let us define an equal-tailed two-sided $(1-\alpha)\%$ credible interval from a combination of two upper one-sided intervals as:

$$\text{two-sided } (1-\alpha)\%\text{CrI} = [\theta^{l*}, \theta^{u*}),$$

where θ^{l*} and θ^{u*} satisfy: $\Pr(\theta < \theta^{l*}|data) = \alpha/2$ and $\Pr(\theta < \theta^{u*}|data) = 1 - \alpha/2$. Note that, in our example of two Normal models, these posterior values are calculated as:

$$\Pr(\theta < \theta^*|data) = \int_{-\infty}^{\theta^*} \pi(\theta|data)d\theta = \frac{\int_{-\infty}^{\theta^*} \Big(f_{Norm}((z-\theta\sqrt{n}), 0, 1) \times \pi(\theta)\Big)d\theta}{\int_{-\infty}^{\infty} \Big(f_{Norm}((z-\theta\sqrt{n}), 0, 1) \times \pi(\theta)\Big)d\theta},$$

where $\pi(\theta)$ is defined as in (2), and the integral in the denominator ensures that the posterior density integrates to one.

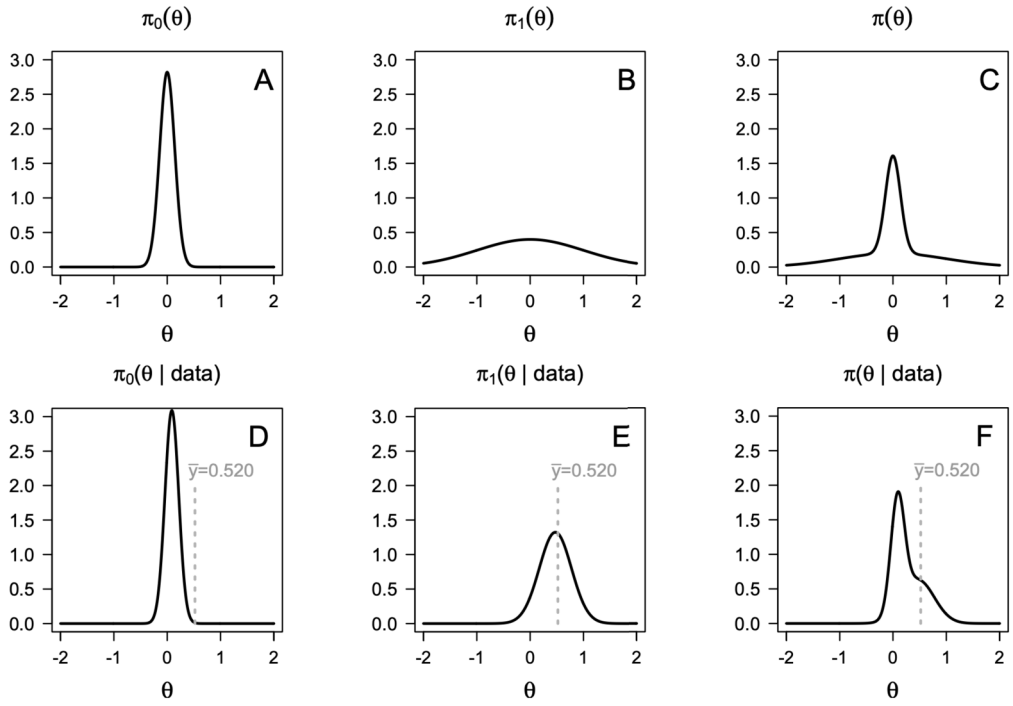

Figure 1: For the "mixture of two normals" example ($g_0 = 0.02$ and $g_1 = 1$), panels A, B, and C, plot the M_0 prior, the M_1 prior, and the mixture-prior, respectively. For data with $\bar{y} = 0.520$ and $n = 10$, panels D, E, and F, plot the M_0 posterior, the M_1 posterior, and the model-averaged posterior, respectively.

Now suppose $g_0 = 0.02$, $g_1 = 1$ and that we observe data for which $\bar{y} = 1.645/\sqrt{n}$ which corresponds to a p-value of $p = 0.05$ when using these data to test against the null hypothesis $\mathrm{H}_0 : \theta < 0$. See Figure 1 which plots priors and posteriors for this scenario with $n = 10$. The lower bound of an upper one-sided $(1 - A)\%$ confidence interval (CI) will be equal to $\mathrm{CI}_A = \bar{y} - \frac{Q_{Norm}(1-A)}{\sqrt{n}}$, where $Q_{Norm}()$ is the Normal quantile function. For instance, for the observed data with $\bar{y} = 1.645/\sqrt{n}$, we have $\mathrm{CI}_{0.10} = (1.645 - 1.282)/\sqrt{n}$, such that an upper one-sided 90% CI will be $= [0.363/\sqrt{n}, \infty)$. An upper one-sided 95% CI for these data will be $[0, \infty)$, since $\mathrm{CI}_{0.05} = 0$. How do these frequentist intervals compare to model-averaged Bayesian credible intervals? While most literature describing the asymptotic agreement of Bayes and frequentist inferences considers the regime of a fixed true parameter value as n increases, for our purposes it is useful to consider the regime of a fixed p-value for a particular point null hypothesis. Consider two observations.

First, setting $\theta^* = 0$ in (5), we see that as n increases (and $p = 0.05$ remains fixed), the corresponding value of α approaches $p = 0.05$: For $n = 10$, we obtain $\alpha = 0.160$, whereas for $n = 10000$, we obtain $\alpha = 0.050$; see how the $\Pr(\theta < \mathrm{CI}_{0.05}|data)$ curve

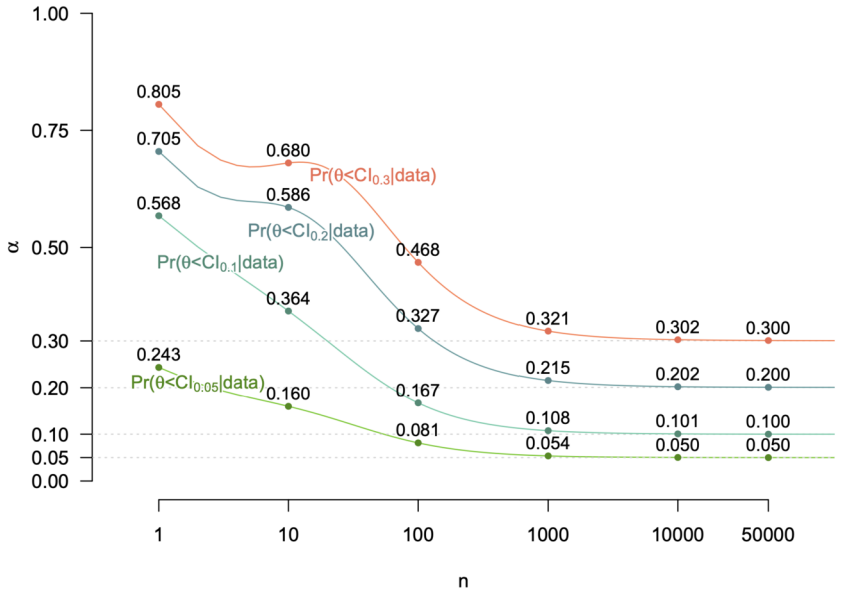

Figure 2: Let CI_A be the lower bound of a frequentist upper one-sided (1-A)% confidence interval. We consider $\Pr(\theta < \text{CI}_A|data) = \alpha$ and data corresponding to (n, p), where n is the sample size and p is the frequentist p-value obtained when testing the data against the null hypothesis $\text{H}_0 : \theta < 0$. While only six specific values of n are highlighted in the plot, the curves are the result of linearly interpolating across 200 different n values equally spaced (on the logarithmic scale) between 1 and 90000. For the normal mixture example with $g_0 = 0.02$ and $g_1 = 1$, and $p = 0.05$, we have $\text{CI}_{0.05} = 0$ and see that, as n increases, α approaches A for $A = 0.05, 0.10, 0.20$ and 0.3. The non-monotonicity of the curves (especially for $A = 0.30$) is notable.

approaches 0.05 as n increases in Figure 2. Second, setting $\theta^* = \text{CI}_A$ in (5), we see that as n increases (and $p = 0.05$ remains fixed), the corresponding value of α approaches A. In Figure 2, we plot values of α corresponding to $A = 0.05, 0.10, 0.20$, and 0.30. One can clearly see that each $\Pr(\theta < \text{CI}_A|data)$ curve tends asymptotically towards A. One can verify this asymptotic behavior by re-expressing posterior expectations arising from the specified prior as posterior expectations arising under an improper uniform prior. For completeness, we give the necessary details in the Supplemental Material (Campbell and Gustafson, 2023).

Based on the asymptotic behavior of the posterior in this example, one might reasonably conclude that, with a sufficiently large sample size, the model-averaged credible interval will approximate the frequentist's confidence interval for any A probability level. However, Wagenmakers and Ly (2021) argue that, in this scenario, "the Jeffreys-Lindley paradox still applies" indicating that there is indeed a conflict between Bayesian and frequentist interpretations of the data.

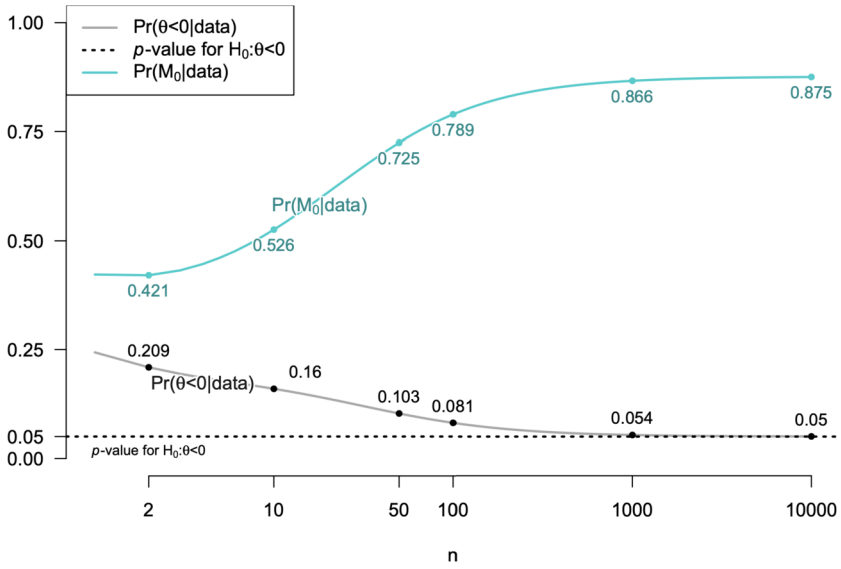

Figure 3: For the normal mixture model example with $g_0 = 0.02$ and $g_1 = 1$, the $\Pr(M_0|data)$ (blue curve) increases towards 0.876 with increasing n, while the value of $\Pr(\theta < 0|data)$ (grey line) approaches 0.05 (dashed black line). While only six specific values of n are highlighted in the plot, the curves are the result of linearly interpolating across 100 different n values equally spaced (on the logarithmic scale) between 1 and 10000.

Wagenmakers and Ly (2021) explain their reasoning as follows. From (4), we calculate $\lim_{n\to\infty}\Pr(M_1|data) = (1 + \sqrt{g_1/g_0})^{-1} = (1 + 1/\sqrt{0.02})^{-1} = 0.124$ and $\lim_{n\to\infty}\Pr(M_0|data) = 0.876$. Therefore, with sufficiently large n, we have that $\Pr(M_1|data) < \Pr(M_0|data)$ regardless of the data (i.e., regardless of the fixed value of $z = \sqrt{n}\bar{y}$); see Figure 3.

In this scenario, model selection (i.e., evaluating the relative values of $\Pr(M_0|data)$ and $\Pr(M_1|data)$) is not addressing the same question as estimation (i.e., evaluating $\Pr(\theta|data)$ to determine which values of θ are *a posteriori* most likely). The posterior density of θ describes one's belief in the probability of different possible values of θ, whereas the posterior model probabilities describe the probability of different data generating processes (DGP) (including the generation of θ). As such, while it is perhaps true that the Jeffreys-Lindley paradox still applies with regards to model selection (i.e., with a sufficiently large sample size and fixed z, the Bayesian will inevitably select M_0) (but do see arguments for and against this in Gray et al. (2023)), the paradox certainly does not apply when it comes to parameter estimation (i.e., with a sufficiently large sample size and fixed z, the Bayesian will inevitably agree with the frequentist when it comes to estimating θ, with their credible interval approximately equal to the frequentist's confidence interval). One way to think about this is to consider the diminishing influence of the prior as the sample size increases and to recall that the confidence inter-

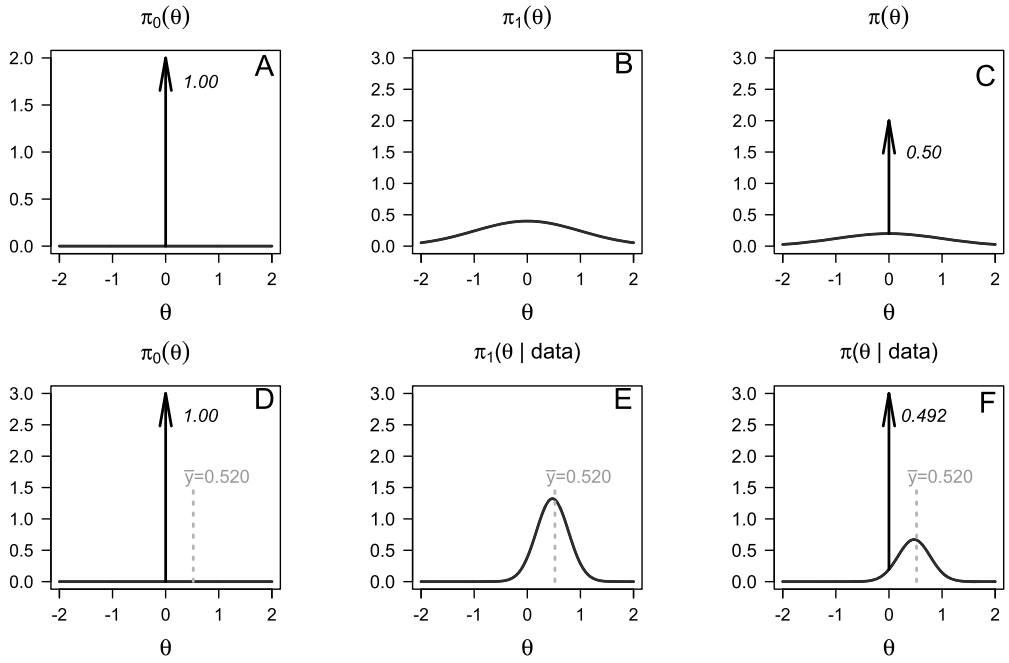

Figure 4: For the "point-null" example, panels A, B, and C, plot the M_0 prior, the M_1 prior, and the mixture-prior, respectively. For data with $\bar{y} = 0.520$ and $n = 10$, panels D, E, and F, plot the M_0 posterior, the M_1 posterior, and the model-averaged posterior, respectively.

val and the credible interval will agree exactly if one specifies the flat (albeit improper) reference prior, $\pi(\theta) \propto 1$; see the worked examples in Held (2020).

In order for the Jeffreys-Lindley paradox to apply to parameter estimation, a point-mass in the prior is required. We consider this situation in the next Section.

3 Parameter estimation with a point null

Consider the same scenario as above but with the null model, M_0, defined as a so-called "point-null" such that the prior density function under M_0 is:

$$\pi_0(\theta) = \delta_0(\theta), \tag{6}$$

where $\delta_0()$ is the Dirac delta function at 0 which can be informally thought of as setting $g_0 = 0$ in (3), or alternatively thought of as a probability density function which is zero everywhere except at 0, where it is infinite. Note that these are merely informal, intuitive interpretations.

We now have that $\Pr(\theta = 0|data) = \Pr(M_0|data)$, or equivalently, $\Pr(\theta \neq 0|data) = \Pr(M_1|data)$. As such, model selection (selecting between M_0 and M_1) and null hy-

pothesis testing (selecting between $\mathrm{H}_0 : \theta = 0$ and $\mathrm{H}_1 : \theta \neq 0$) are equivalent in this scenario.

With the "point-null" prior for M_0 as defined in (6), and with $g_1 = 1$, as defined previously, the "mixture" prior, $\pi(\theta)$, is recognizable as a "spike-and-slab" prior (see van den Bergh et al. (2021)) and the Bayes factor is equal to:

$$\mathrm{BF}_{01} = \sqrt{1+n} \times \exp\Big(\frac{-nz^2}{2(1+n)}\Big).$$

The posterior density is nonatomic with a spike (i.e., a discontinuity with infinite density) at 0:

$$\pi(\theta|data) = \Pr(M_0|data)\delta_0(\theta) + \Pr(M_1|data) f_{Normal}\Big(\theta, \frac{z}{\sqrt{n}(\frac{1}{n}+1)}, \frac{1}{1+n}\Big),$$

where the posterior model probabilities, $\Pr(M_0|data)$ and $\Pr(M_1|data)$, can be calculated from the Bayes factor as in (4).

Returning to our hypothetical data with $z = 1.645$, we see that for $\theta^* = 0$, as n increases, α (such that $\Pr(\theta < \theta^*|data) = \alpha$) does not approach $p = 0.05$ and instead approaches 0: For $n = 10$, we obtain $\alpha = 0.03$, and for $n = 1000$, we obtain $\alpha = 0.005$; see trajectory of the grey curve in Figure 5. What's more, as n increases and $\bar{y} = 1.645/\sqrt{n}$ remains fixed, the posterior probability on the "spike" at 0 increases towards infinity such that: $\lim_{n\to\infty}\Pr(M_0|data) = 1$; as famously emphasized by Lindley (1957) and originally demonstrated by Jeffreys (1935).

Perhaps even more puzzling is that, for fixed $\alpha = 0.05$, there is simply no corresponding value of θ^* (such that $\alpha = \Pr(\theta < \theta^*|data)$) for any $n > 2$. For $n = 2$ we can define $\theta^* = -0.0163$, such that $\Pr(\theta < -0.0163|data) = 0.05$. However, for $n = 3$, a precise value of θ^* cannot be defined since, due to the discontinuity in the posterior, we have: $\Pr(\theta < 0|data) = 0.045 < \alpha$, and $\Pr(\theta \leq 0|data) = 0.465 > \alpha$. For $n = 10$ the gap is even wider: $\Pr(\theta < 0|data) = 0.030 < \alpha$ and $\Pr(\theta \leq 0|data) = 0.522 > \alpha$. Figure 5 plots these numbers for increasing values of n. As a consequence, it is no longer the case that, with a sufficiently large sample size, a Bayesian's credible interval will approximate a frequentist's confidence interval. In fact, for certain values of α and n, calculating a credible interval is not even possible.

In general, determining a specific value of θ^* for a given value of α (such that $\alpha = \Pr(\theta < \theta^*|data)$) is only possible for values of α outside of the "incredibility interval":

$$\Big[\Big(\Pr(\theta < 0|data, M_1)\Pr(M_1|data)\Big), \Big(\Pr(\theta < 0|data, M_1)\Pr(M_1|data) + \Pr(M_0|data)\Big)\Big].$$

The bounds of the "incredibility interval" are the limits of the "jump" in the cumulative distribution function of the posterior, i.e., the values ranging between $\Pr(\theta < 0|\mathrm{data})$ and $\Pr(\theta \leq 0|\mathrm{data})$. In Figure 6, we plot the cumulative distribution function of the posterior for hypothetical data with $z = 1.645$ and $n = 10$. In this situation, the "incredibility interval" equals $[\Pr(\theta < 0|data), \Pr(\theta \leq 0|data)] = [0.03, 0.522]$. In Figure 5, the lower grey curve corresponds to the lower bound of the incredibil-

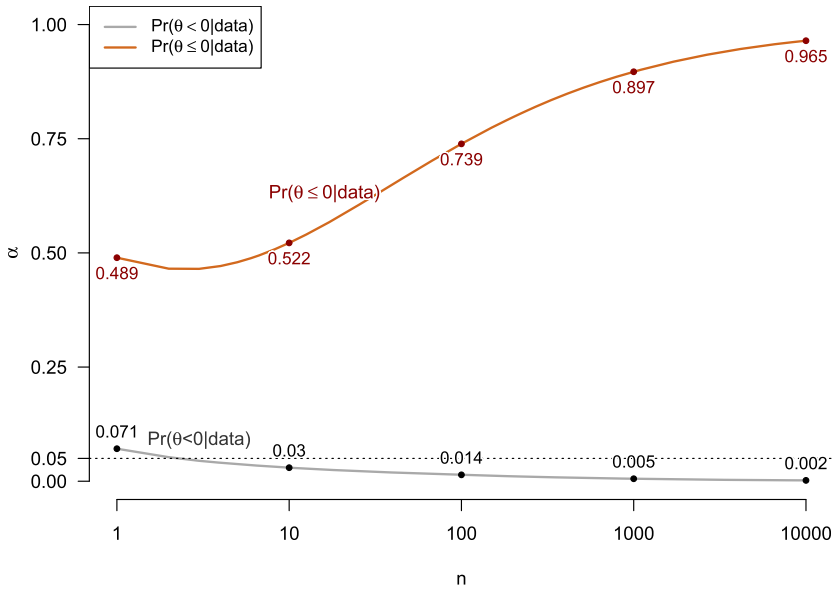

Figure 5: For the hypothetical data with $z = 1.645$, as n increases along the horizontal axis, values of α such that $\Pr(\theta < 0|data) = \alpha$ (grey line) and $\Pr(\theta \leq 0|data) = \alpha$ (red line) are plotted on the vertical axis. While only five specific values of n are highlighted in the plot, the curves are the result of linearly interpolating across 100 different n values equally spaced (on the logarithmic scale) between 1 and 10000.

ity interval and the upper red curve corresponds to the upper bound. Notably, since $\lim_{n\to\infty}\Pr(M_0|data) = 1$ and $\lim_{n\to\infty}\Pr(M_1|data) = 0$, the width of the incredibility interval increases as n increases. As a result, determining a precisely α-level value of θ^* such that $\alpha = \Pr(\theta < \theta^*|data)$, becomes increasingly impossible as n grows large. This is true regardless of the data; see Figure 7 for values of the lower bound obtained with data where $\bar{y} = 2.575/\sqrt{n}$ (data for which one obtains a p-value of $p = 0.005$ when testing against $\mathrm{H}_0 : \theta < \theta_0$).

When α is inside the incredibility interval, there remains an unconventional way for defining a $(1-\alpha)\%$ credible interval. In order to establish a correct value for θ^* such that $\Pr(\theta < \theta^*|data) = \alpha$ (over repeated samples) one defines θ^* stochastically such that

$$\theta^* = \begin{cases} 0, & \text{with probability } \gamma; \text{ and} \\ 0+\epsilon, & \text{with probability } 1-\gamma, \end{cases} \tag{7}$$

where:

$$\gamma = \frac{\alpha - \Pr(\theta \leq 0|data)}{\Pr(\theta < 0|data) - \Pr(\theta \leq 0|data)},$$

and ϵ is an arbitrarily small number.

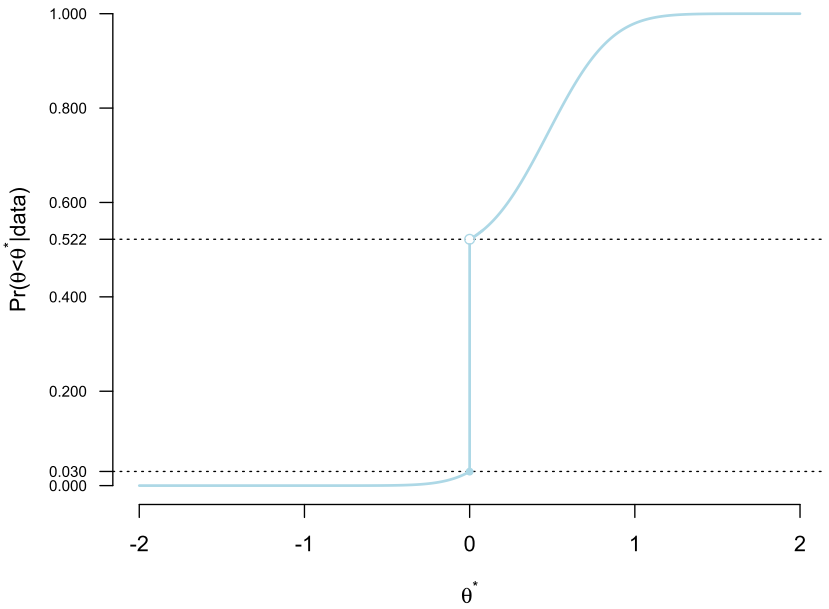

Figure 6: For the hypothetical data with $z = 1.645$ and $n = 10$, the plotted line corresponds to the cumulative distribution function of the posterior (i.e., $\Pr(\theta < \theta^*|data)$) for increasing values of θ^*.

Returning to our example data with $\bar{y} = 1.645/\sqrt{n}$, we note that, for $n = 10$, $\Pr(\theta < 0|data) = 0.030$ and $\Pr(\theta \leq 0|data) = 0.522$. As such, for $\alpha = 0.05$ (which is inside the incredibility interval of [0.030, 0.522]), we define θ^* as:

$$\theta^* = \begin{cases} 0, & \text{with probability } \gamma = 0.959; \text{ and} \\ 0 + \epsilon, & \text{with probability } (1 - \gamma) = 0.041. \end{cases}$$

Defining θ^* in this way will guarantee that $\Pr(\theta < \theta^*|data) = 0.05$. One way to think about this is to consider the various values of θ that, over a researcher's lifetime give rise to the various datasets they analyse. Across all of these studies, the average posterior probability content of the $[\theta^*, \infty)$ interval will be 0.95. Thinking about hypothetical replications in this way has an admittedly frequentist character. However, these are replications across studies arising from different parameter values. If the model-averaged prior does in fact correspond to the true data generating mechanism, we can be assured that, amongst all of the researcher's studies for which $z = 1.645$, 95% of these were the result of a θ value from inside of their interval. Furthermore, since this is true for any arbitrary value of z and any arbitrary value of α, then we have that $\Pr\big(\theta_j \in [\theta^*, \infty)\big|z_j\big) = 1 - \alpha$, where θ_j and z_j are values obtained from a joint draw from the amalgamation of the prior and statistical model (i.e., the data generating mechanism).

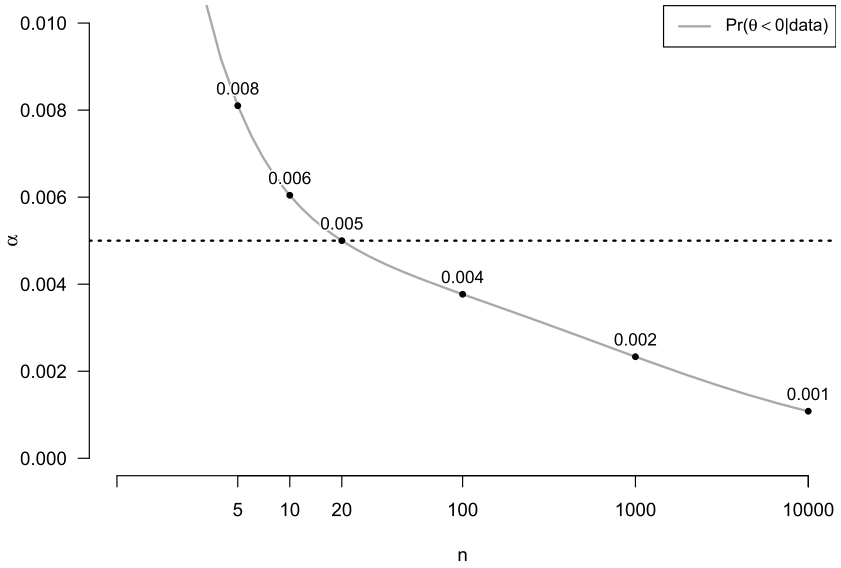

Figure 7: With data where $\bar{y} = 2.575/\sqrt{n}$, as n increases, the lower bound of the incredibility interval (the solid line) decreases towards zero. As a consequence, determining a value of θ^* such that $\Pr(\theta < \theta^* | data) = \alpha$, when $\alpha = 0.005$ (the dotted line) is only possible for $n < 20$. While only six specific values of n are highlighted in the plot, the curve is the result of linearly interpolating across 100 different n values equally spaced (on the logarithmic scale) between 1 and 10000.

As another example, suppose $n = 100$ and $\bar{y} = 2.054/\sqrt{n} = 0.2054$ which corresponds to a p-value of $p = 0.04$ when using the data to test against the null hypothesis $\mathrm{H}_0 : \theta = 0$, and a p-value of $p = 0.02$ when using the data to test against the null hypothesis $\mathrm{H}_0 : \theta < 0$. One can easily calculate an upper one-sided frequentist 95% confidence interval for these data equal to: $[\bar{y} - 1.645/\sqrt{n}, \infty) = [0.040, \infty)$, which clearly excludes 0. However, one cannot calculate an upper one-sided 95% credible interval since $\alpha = 0.05$ is within the incredibility interval for this data: [0.009, 0.564]. The closest one can do is to calculate an upper one-sided 99.1% credible equal to: $[0, \infty)$ which includes 0, or calculate an upper one-sided 43.6% credible interval equal to $(0, \infty)$ which excludes 0. The only way to define an upper one-sided interval with exactly 95% probability of including the true value of θ (over repeated samples) is to do so stochastically as equal to: $[\theta^*, \infty)$, where $\theta^* = 0$ with probability $\gamma = (0.050 - 0.564)/(0.009 - 0.564) = 0.926$, and $\theta^* = 0 + \epsilon$ with probability $1 - \gamma = 0.074$.

We are not seriously suggesting that researchers define credible intervals in this bizarre stochastic way. We simply wish to demonstrate that this is the only way one can correctly define the credible interval from a posterior with point masses. When model-averaged posteriors involve point-null models, credible intervals must therefore be approached and interpreted with the utmost caution. The issue only gets thornier as the sample size increases.

For a very very large n it is possible that both $\alpha/2$ and $(1-\alpha/2)$ are within the incredibility interval. In this case, the equal-tailed two-sided $(1-\alpha)\%$ credible interval must be defined in an even more bizarre way. When both $\alpha/2$ and $(1-\alpha/2)$ are in the incredibility interval, the credible interval must be defined stochastically as either a single point or as an entirely empty interval:

$$(1-\alpha)\%\text{CrI} = \begin{cases} [0], & \text{with probability } \psi; \text{ and} \\ \emptyset, & \text{with probability } (1-\psi), \end{cases} \tag{8}$$

where

$$\psi = \frac{\Pr(\theta = 0|data) - \alpha}{2 \times \Pr(\theta = 0|data) - 1}.$$

To be clear, the "stochastic credible interval" is not defined in (7) and (8) to ensure that it has a certain (asymptotic) coverage. Rather it is defined in the only possible way such that (over repeated samples) the boundaries of the interval contain the correct amount of posterior mass (as required by the definition in (5)). As such, it may not be immediately obvious that, when we look at the asymptotic behavior of these stochastic credible intervals, we see that the Jeffreys-Lindley paradox reduces the data to be entirely inconsequential (at least when assuming a fixed p-value). Indeed, as n increases, both γ and ψ approach $1-\alpha$ since:

$$\begin{aligned} \lim_{n\to\infty}\gamma &= \lim_{n\to\infty}\Big(\frac{\alpha - \Pr(\theta \leq \theta_0|data)}{\Pr(\theta < \theta_0|data) - \Pr(\theta \leq \theta_0|data)}\Big) \\ &= \Big(\frac{\alpha - 1}{-1}\Big) \\ &= 1-\alpha, \end{aligned}$$

and:

$$\begin{aligned} \lim_{n\to\infty}\psi &= \lim_{n\to\infty}\Big(\frac{\Pr(\theta = 0|data) - \alpha}{2 \times \Pr(\theta = 0|data) - 1}\Big) \\ &= \Big(\frac{1-\alpha}{2-1}\Big) \\ &= 1-\alpha. \end{aligned}$$

Therefore, for sufficiently large n and z remaining constant, the probability that one will exclude 0 from a $(1-\alpha)\%$credible interval will equal α regardless of the data; see Figure 8. While this may strike one as paradoxical, it is entirely congruent with the widely-known consequence of the Jeffreys-Lindley paradox: As n increases and z is fixed, the probability of selecting M_0 will go to 1.

4 Conclusion

We demonstrated that when one of the two models under consideration is a point-null model, not only can a model-averaged credible interval be rather different than

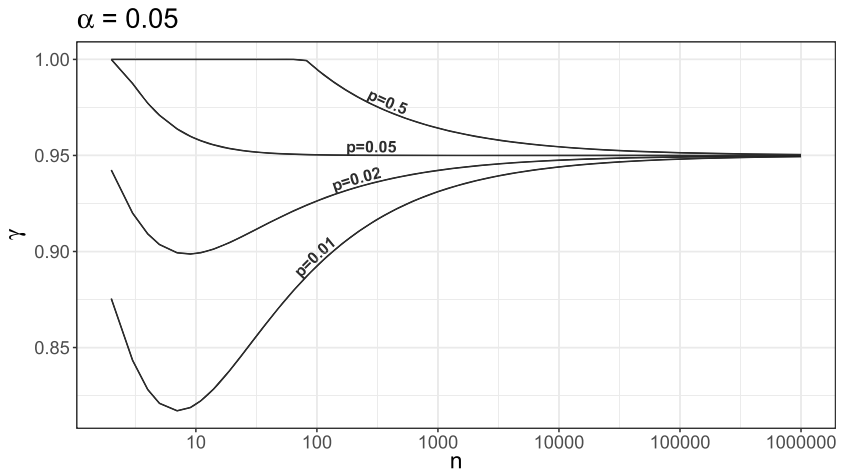

Figure 8: Each line corresponds to observing data corresponding to a p-value of p when testing against $\mathrm{H}_0 : \theta < 0$.

the frequentist confidence interval, oftentimes it will be simply undefined (at least in a conventional sense). As a consequence, it may be tempting to compare (e.g., using the Bayes factor) two *a priori* probable models, M_0 and M_1, for the purpose of model selection, but then simply report the uncertainty about θ, conditional on M_1 being unquestionably true. We caution that this strategy, while seemingly straightforward, will lead to unavoidable inconsistencies between one's priors and posteriors. Campbell and Gustafson (2022) explain in detail why disregarding M_0 "for the purpose of parameter estimation" (Wagenmakers and Gronau, 2020) is inadvisable; see also Tendeiro and Kiers (2019). Frequentists encounter related issues when it comes to obtaining post-selection uncertainty intervals with regularization approaches such as the lasso; see Lu et al. (2017) and the references therein.

Some researchers may be happy to avoid model selection entirely and may see no reason to entertain point-null priors (e.g., Gelman and Rubin (1995): "realistic prior distributions in social science do not have a mass of probability at zero" [...] "we believe model selection to be relatively unimportant compared to the task of constructing realistic models that agree with both theory and data."). However, if researchers truly believe that there is a non-zero prior probability that the parameter of interest is precisely zero (and this prior probability is equal to the value assigned to $\Pr(M_0)$), Bayesian testing with a point-null will be optimal in the sense of minimizing the expected loss (with respect to a joint distribution of the data and parameters); see Berger (1985). These researchers should be aware that, while perhaps optimal, Bayesian testing with a point-null can lead to rather unexpected asymptotic behavior. There will still be credible intervals; it is just that, as a consequence of the discontinuity in the model averaged posterior, certain specific credible intervals do not exist. Some researchers might therefore wish to explore alternative means of conveying the uncertainty surrounding the parameter of interest (e.g., Wagenmakers et al. (2022), Rice and Ye (2022)).

One limitation of this work is that we only considered univariate models where one wishes to define a credible interval for a single parameter of interest. However, the ideas we discussed also apply to multivariate settings where one wishes to define credible sets and where there may be several different nested models under consideration. For instance, researchers using Bayes factors in multiple regression models (Rouder and Morey, 2012) should be aware that it may be impossible to define certain model-averaged credible intervals/sets for the regression coefficients.

A second limitation is that we did not consider how the undefinability of specific credible intervals will also occur in discrete parameter models. In such cases, specific confidence intervals will also be undefined (Tingley and Li, 1993; Berger, 1985) (it is impossible to define a continuous interval on the real line as the parameter can only have countably many, discrete values), so while they may both be puzzled, Bayesians and frequentists should at least agree in their inability to define an uncertainty interval! Instead of uncertainty intervals, one should consider uncertainty sets in such a scenario. This distinction is at the core of the issue that arises when model averaging with a point-null prior and the stochastic credible intervals that include both a discrete value (i.e., zero) and a continuous interval.

Finally, we note that the consequences of the Jeffreys-Lindley paradox on model selection (and null hypothesis testing) are often understood as "intuitive" and not necessarily unfavorable: When sample sizes are very large, researchers might indeed prefer to sacrifice some power in order to lower the probability of a type I error, a trade-off that occurs necessarily when testing a point-null hypothesis with the Bayes factor; see Pericchi and Pereira (2016) and Wagenmakers and Ly (2021). Indeed, the benefits of such a trade-off are routinely discussed by frequentists and Bayesians alike (e.g., Leamer (1978): "from every reasonable viewpoint the significance level should be a decreasing function of sample size"; and recently, Wulff and Taylor (2023): "From a Neyman-Pearson perspective, it is logical that α should be a decreasing function of the sample size."). However, the consequences of the Jeffreys-Lindley paradox on parameter estimation –specifically with regards to model-averaged credible intervals and the inability to define these for certain probability levels– were previously less well understood, and certainly strike us as less intuitive.

Supplementary Material

Supplemental Material. Defining a Credible Interval Is Not Always Possible with "Point-Null" Priors: A Lesser-Known Correlate of the Jeffreys-Lindley Paradox
(DOI: 10.1214/23-BA1397SUPP; .pdf).

References

Albers, C. J., Kiers, H. A. and van Ravenzwaaij, D. (2018), 'Credible confidence: A pragmatic view on the frequentist vs Bayesian debate', *Collabra: Psychology* **4**(1). 1

Berger, J. O. (1985), *Statistical decision theory and Bayesian analysis*, Springer Sci-

ence & Business Media. MR0804611. doi: https://doi.org/10.1007/978-1-4757-4286-2. 13, 14

Campbell, H. and Gustafson, P. (2022), 'Bayes factors and posterior estimation: Two sides of the very same coin', *arXiv preprint arXiv:2204.06054*. 1, 2, 13

Campbell, H. and Gustafson, P. (2023). 'Supplemental Material. Defining a Credible Interval Is Not Always Possible with "Point-Null" Priors: A Lesser-Known Correlate of the Jeffreys-Lindley Paradox.' *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1397SUPP. 5

Casella, G. and Berger, R. L. (1987), 'Reconciling Bayesian and frequentist evidence in the one-sided testing problem', *Journal of the American Statistical Association* **82**(397), 106–111. MR0883339. doi: https://doi.org/10.1080/01621459.1987.10478396. 2

Datta, G. S. and Ghosh, J. K. (1995), 'On priors providing frequentist validity for Bayesian inference', *Biometrika* **82**(1), 37–45. MR1332838. doi: https://doi.org/10.2307/2337625. 2

Gelman, A. and Rubin, D. B. (1995), 'Avoiding model selection in Bayesian social research', *Sociological Methodology* **25**, 165–173. doi: https://doi.org/10.2307/271064. 13

Gray, J., Cherry, J. L., Wagenmakers, E.-J. and Ly, A. (2023), 'The Jeffreys–Lindley paradox: an exchange', *Archive for History of Exact Sciences* **77**, 443–449. MR4604374. doi: https://doi.org/10.1007/s00407-023-00310-4. 6

Greenland, S. and Poole, C. (2013), 'Living with p-values: Resurrecting a Bayesian perspective on frequentist statistics', *Epidemiology* **24**(1), 62–68. 2

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. et al. (2023), 'A review of applications of the Bayes factor in psychological research', *Psychological Methods* **28**(3), 558–579. 1

Held, L. (2020), Bayesian tail probabilities for decision making, *in* 'Bayesian Methods in Pharmaceutical Research', CRC Press Taylor & Francis Group, pp. 53–73. MR4599165. 2, 7

Jeffreys, H. (1935), Some tests of significance, treated by the theory of probability, *in* 'Mathematical proceedings of the Cambridge philosophical society', Vol. 31, Cambridge University Press, pp. 203–222. 8

Keysers, C., Gazzola, V. and Wagenmakers, E.-J. (2020), 'Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence', *Nature Neuroscience* **23**(7), 788–799. doi: https://doi.org/10.1038/s41593-020-0660-4. 1

Leamer, E. E. (1978), *Specification searches: Ad hoc inference with nonexperimental data*, Vol. 53, John Wiley & Sons Incorporated. MR0471118. 14

Lindley, D. V. (1957), 'A statistical paradox', *Biometrika* **44**(1/2), 187–192. MR0087273. doi: https://doi.org/10.1093/biomet/44.1-2.187. 8

Lu, S., Liu, Y., Yin, L. and Zhang, K. (2017), 'Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* pp. 589–611. MR3611761. doi: https://doi.org/10.1111/rssb.12184. 13

Pericchi, L. and Pereira, C. (2016), 'Adaptative significance levels using optimal decision rules: balancing by weighting the error probabilities', *Brazilian Journal of Probability and Statistics* **30**(1), 70–90. MR3453515. doi: https://doi.org/10.1214/14-BJPS257. 14

Rice, K. and Ye, L. (2022), 'Expressing regret: a unified view of credible intervals', *The American Statistician* **76**(3), 248–256. MR4453527. doi: https://doi.org/10.1080/00031305.2022.2039764. 13

Rouder, J. N. and Morey, R. D. (2012), 'Default Bayes factors for model selection in regression', *Multivariate Behavioral Research* **47**(6), 877–903. doi: https://doi.org/10.1080/00273171.2012.734737. 14

Tendeiro, J. N. and Kiers, H. A. (2019), 'A review of issues about null hypothesis Bayesian testing.', *Psychological Methods* **24**(6), 774. doi: https://doi.org/10.1037/met0000221. 13

Tingley, M. and Li, C. (1993), 'A note on obtaining confidence intervals for discrete parameters', *The American Statistician* **47**(1), 20–23. MR1207889. doi: https://doi.org/10.2307/2684776. 14

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N. and Wagenmakers, E.-J. (2021), 'A cautionary note on estimating effect size', *Advances in Methods and Practices in Psychological Science* **4**(1). doi: https://doi.org/10.1177/2515245921992035. 8

Wagenmakers, E.-J. and Gronau, Q. F. (2020), 'Overwhelming evidence for vaccine efficacy in the Pfizer trial: An interim Bayesian analysis', PsyArXiv. doi: https://doi.org/10.31234/osf.io/fs562. 13

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F. and Etz, A. (2022), 'The support interval', *Erkenntnis* **87**, 589–601. MR4396731. doi: https://doi.org/10.1007/s10670-019-00209-z. 13

Wagenmakers, E.-J. and Ly, A. (2021), 'History and nature of the Jeffreys-Lindley paradox', *arXiv preprint arXiv:2111.10191*. MR4532732. doi: https://doi.org/10.1007/s00407-022-00298-3. 2, 5, 6, 14

Wasserstein, R. L. and Lazar, N. A. (2016), 'The ASA statement on p-values: context, process, and purpose', *The American Statistician* **70**(2), 129–133. MR3511040. doi: https://doi.org/10.1080/00031305.2016.1154108. 1

Wulff, J. N. and Taylor, L. (2023), 'How and why alpha should depend on sample size: A Bayesian-frequentist compromise for significance testing'. doi: https://doi.org/10.31234/osf.io/3cbh7. 14