Defining a credible interval is not always possible with "point-null" priors: A lesser-known consequence of the Jeffreys-Lindley paradox

Harlan Campbell and Paul Gustafson
Department of Statistics, University of British Columbia

September 30, 2022

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 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 undefinable. This is a lesser-known consequence 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, several Bayesian tests using Bayes factors have been proposed as alternatives to frequentist hypothesis testing; see Heck et al. (2022) for a recent 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 & Lazar (2016).

Campbell & Gustafson (2022) consider how Bayesian testing and estimation can be done in a complimentary manner and conclude 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.

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 & Berger (1987), Greenland & Poole (2013), Held et al. (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 a 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 & 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 & Ly (2021) in their discussion of the Jeffreys-Lindley paradox. In Section 3, we extend this example to consider the consequences 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 Campbell & 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

$$\pi_1(\theta) = f_{Normal}(\theta, 0, g_1), \tag{4}$$

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 $i = 1, ..., n$; 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:

$$\Pr(data|\theta) = \prod_{i=1}^{n} f_{Normal}(y_i, \theta, 1). \tag{5}$$

Then the Bayes factor is:

$$\mathrm{BF}_{01} = \sqrt{\frac{1+ng_1}{1+ng_0}} \times \exp\Big(\frac{(g_0 - g_1)nz^2}{2(1+ng_0)(1+ng_1)}\Big), \tag{6}$$

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

$$\Pr(M_0|data) = \frac{\Pr(M_0)}{\Pr(M_1)/\mathrm{BF}_{01} + \Pr(M_0)} \quad \text{and} \quad \Pr(M_1|data) = 1 - \Pr(M_0|data). \tag{7}$$

Finally, the model specific posteriors are defined as:

$$\pi_j(\theta|data) = f_{Normal}\Big(\theta, \frac{zg_j}{\sqrt{n}(\frac{1}{n} + g_j)}, \frac{g_j}{1+g_jn}\Big), \tag{8}$$

for $j = 0, 1$.

Having established all the components of equation (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:

$$\text{one-sided } (1-\alpha)\%\mathrm{CrI} = [\theta^*, \infty), \tag{9}$$

where θ^* satisfies the following equality:

$$\Pr(\theta < \theta^*|data) = \alpha. \tag{10}$$

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

$$\text{two-sided } (1-\alpha)\%\mathrm{CrI} = [\theta^{l*}, \theta^{u*}), \tag{11}$$

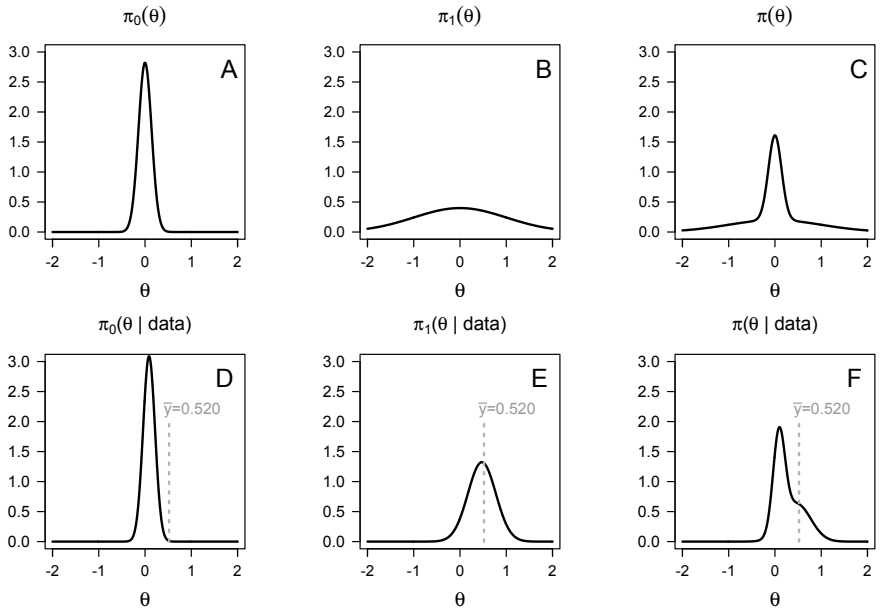

Figure 1: For the "mixture of two normals" example, panels A, B, and C, plot the M_0 prior, the M_1 prior, and the mixture-prior, respectively. For data with $\bar{y} = 0.520$ and $n = 10$, panels D, E, and F, plot the M_0 posterior, the M_1 posterior, and the model-averaged posterior, respectively.

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 equation (2).

To illustrate, let $g_0 = 0.02$, $g_1 = 1$ and suppose 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 $n = 10$. A frequentist upper one-sided 95% confidence interval for these data will be: $[0, \infty)$. A 90% equal tailed two-sided confidence interval will be: $[0, 2 \times 1.645/\sqrt{n})$. How do these frequentist intervals compare to model-averaged Bayesian credible intervals? Consider two observations on this.

First, setting $\alpha = 0.05$ in equation (10), we see that as n increases, θ^* approaches 0: For $n = 10$, we obtain $\theta^* = -0.1039$, whereas for $n = 10000$, we obtain $\theta^* = -0.0001$; see lower panel of Figure 2. For $n = 10$, a 90% equal tailed two-sided confidence interval will be [0,1.0404) and a 90% equal-tailed credible interval will be

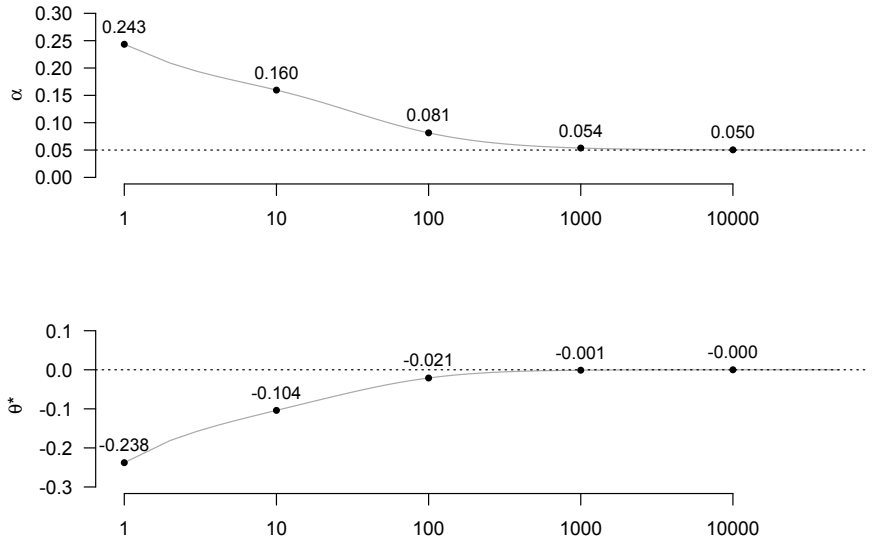

Figure 2: We consider $\Pr(\theta < \theta^*|data) = \alpha$ and data corresponding to $(n,\ p)$, where n is the sample size and p is the frequentist p-value obtained when testing the data against the null hypothesis $\mathrm{H}_0 : \theta < \theta_0$. For the normal mixture example with $g_0 = 0.1$ and $g_1 = 1$, and $p = 0.05$ for $\theta^* = \theta_0 = 0$, we see that, as n increases, α approaches p (upper panel). For $\alpha = p = 0.05$, as n increases, θ^* approaches θ_0 (lower panel).

$[-0.1039, 0.8501)$. For for $n = 10000$, a 90% equal tailed two-sided confidence interval will be [0,0.0329) and a 90% equal-tailed credible interval will be $[-0.0001, 0.0328)$.

Secondly, setting $\theta^* = 0$, we see that as n increases, 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 upper panel of Figure 2. This asymptotic behaviour also holds for arbitrary values of θ^*. As n increases, the value of $\Pr(\theta < \theta^*|data)$ will approach $\int_{-\infty}^{\sqrt{n}(\bar{y}-\theta^*)} f_{Norm}(x, 0, 1)dx$, which is equal to the frequentist p-value obtained when testing the data against the null hypothesis $\mathrm{H}_0 : \theta < \theta^*$. Note how the solid and dashed curves approach one another as n increases in Figure 3, for $\theta^* = 0.01$ (green curves), for $\theta^* = 0.10$ (grey curves), and for $\theta^* = 0.35$ (red curves).

Based on the asymptotic behaviour 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 α probability level. However, Wagenmakers & 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.

Wagenmakers & Ly (2021) explain their reasoning as follows. From equation (7),

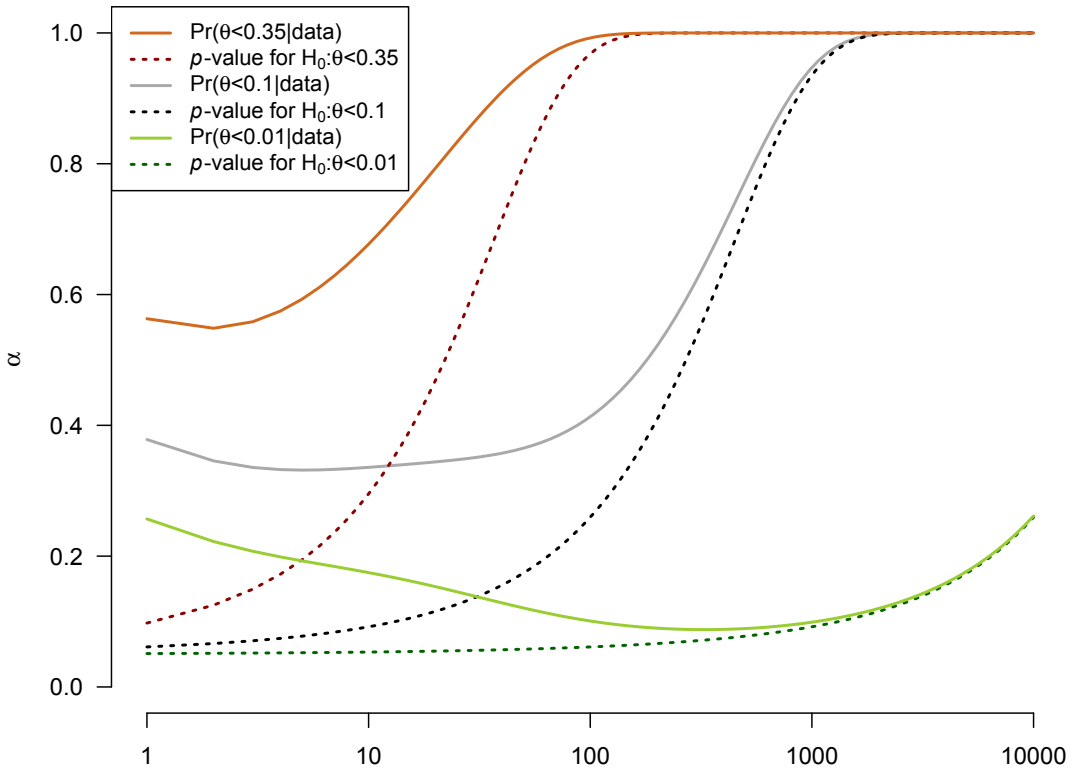

Figure 3: Let $\bar{y} = 1.645/\sqrt{n}$ and let $g_0 = 0.1$. As n increases, the value of $\Pr(\theta < \theta^*|data)$ (solid line) and $\int_{-\infty}^{\sqrt{n}(\bar{y}-\theta^*)} f_{Norm}(x, 0, 1)dx$ (dashed line) approach one another; for $\theta^* = 0.01$ (green curves), for $\theta^* = 0.10$ (grey curves), and for $\theta^* = 0.35$ (red curves).

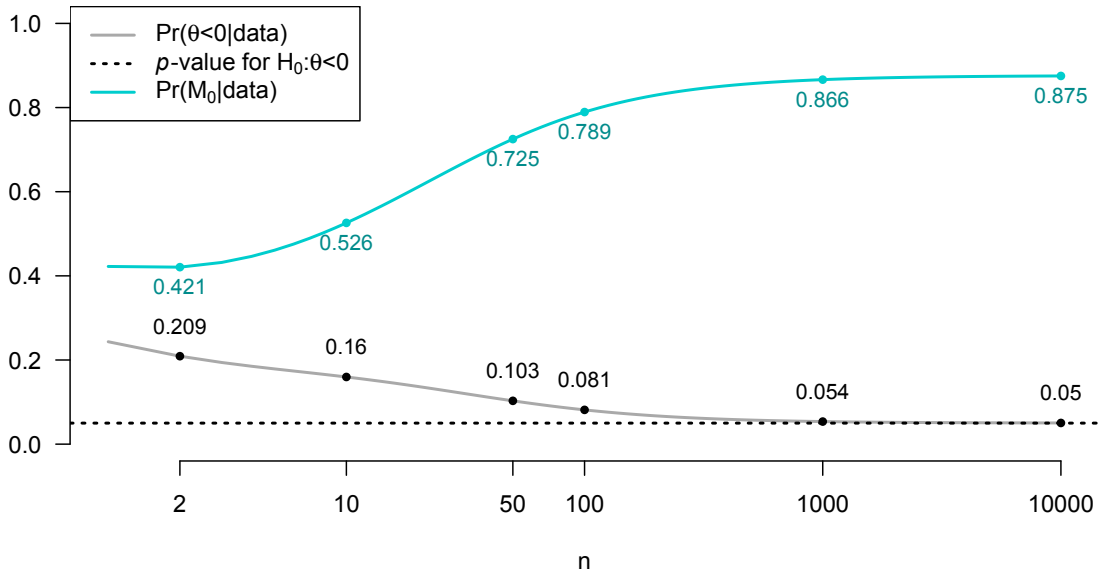

Figure 4: For the normal mixture model example with $g_0 = 0.02$ and $g_1 = 1$, the $\Pr(M_1|data)$ (blue curve) increases towards 0.876 with increasing n, while the value of $\Pr(\theta < 0|data)$ (grey line) approaches 0.05 (dashed black line).

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 4.

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 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), the paradox 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). 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.

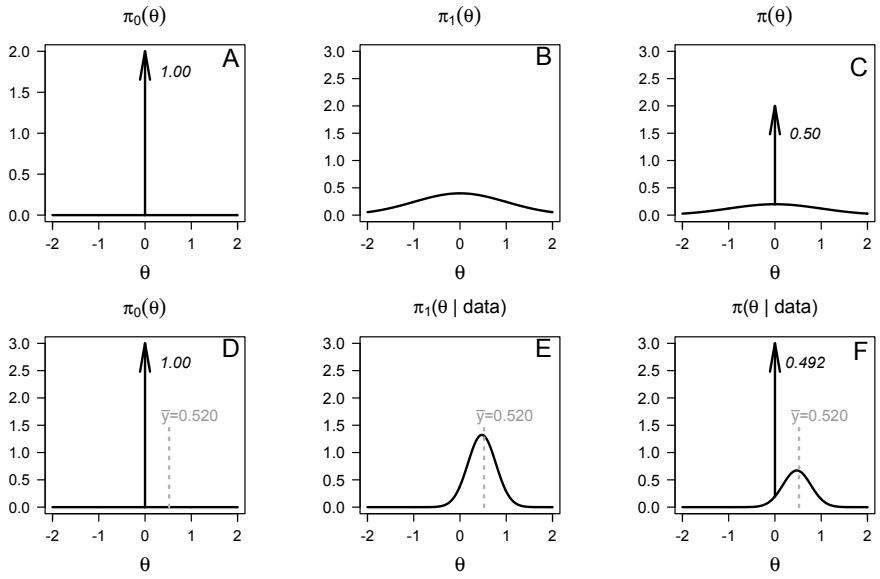

Figure 5: For the "point-null" example, panels A, B, and C, plot the M_0 prior, the M_1 prior, and the mixture-prior, respectively. For data with $\bar{y} = 0.520$ and $n = 10$, panels D, E, and F, plot the M_0 posterior, the M_1 posterior, and the model-averaged posterior, respectively.

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:

$$\pi_0(\theta) = \delta_0(\theta), \tag{12}$$

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

Note 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 hypothesis 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 (12), 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:

$$\mathrm{BF}_{01} = \sqrt{1+n} \times \exp\Big(\frac{-nz^2}{2(1+n)}\Big), \tag{13}$$

The posterior density is nonatomic with a spike (i.e., a discontinuity with infinite

density) at 0:

$$\pi(\theta|data) = \Pr(M_0|data)\delta_0(\theta) + \Pr(M_1|data) f_{Normal}\Big(\theta, \frac{z}{\sqrt{n}(\frac{1}{n}+1)}, \frac{1}{1+n}\Big), \quad (14)$$

where the posterior model probabilities, $\Pr(M_0|data)$ and $\Pr(M_1|data)$, can be calculated from the Bayes factor as in equation (7).

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 6. Whatsmore, 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 recognized by Lindley (1957).

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 6 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":

$$\Big[\Big(\Pr(\theta < 0|data, M_1)\Pr(M_1|data)\Big), \Big(\Pr(\theta < 0|data, M_1)\Pr(M_1|data) + \Pr(M_0|data)\Big)\Big]. \quad (15)$$

In Figure 6, the lower grey curve corresponds to the lower bound of the incredibility 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 a unconventional way for defining a $(1-\alpha)\%$ credible interval. In order to establish a correct value for θ^*

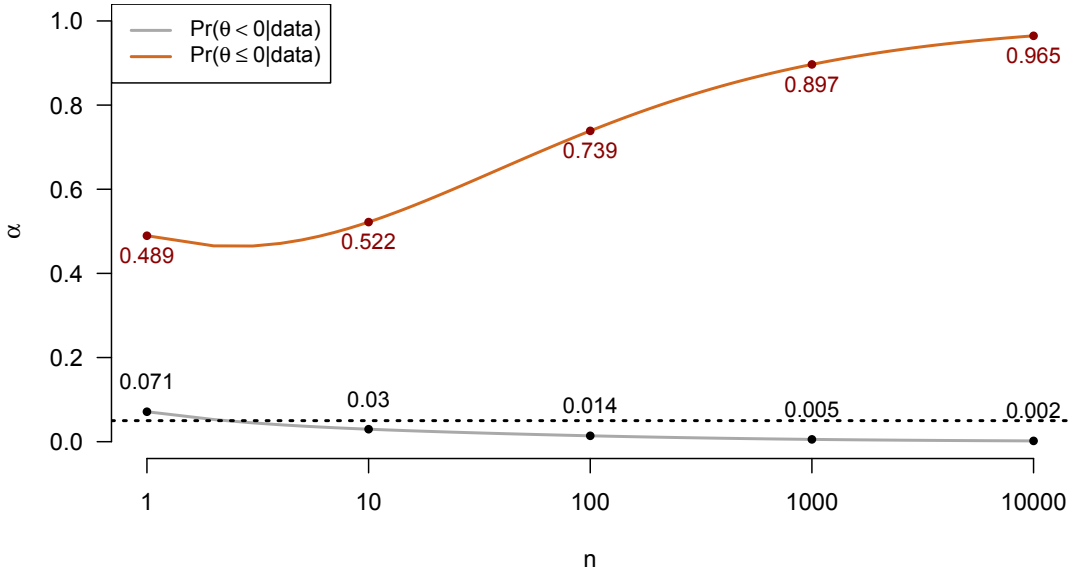

Figure 6: For the hypothetical data with $z = 1.645$, as n increases along the horizontal axis, values of α such that $\Pr(\theta < 0|data) = \alpha$ (grey line) and $\Pr(\theta \leq 0|data) = \alpha$ (red line) are plotted on the vertical axis.

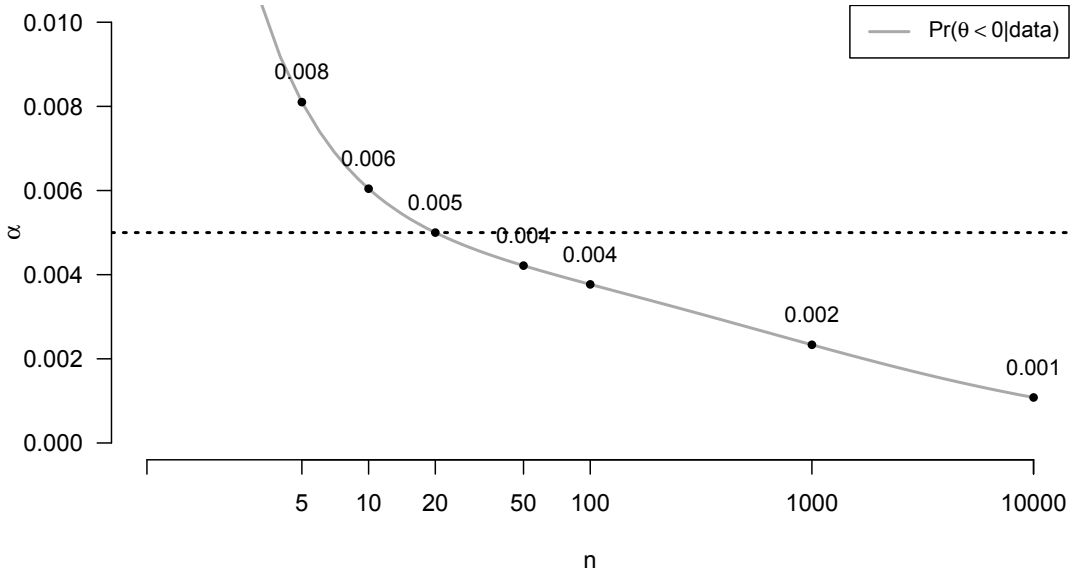

Figure 7: With data where $\bar{y} = 2.575/\sqrt{n}$, as n increases, the lower bound of the incredibility interval (the solid line) decreases towards zero. As a consequence, determining a value of θ^* such that $\Pr(\theta < \theta^*|data) = \alpha$, when $\alpha = 0.005$ (the dotted line) is only possible for $n < 20$.

such that $\Pr(\theta < \theta^* | data) = \alpha$ (over repeated samples) one defines θ^* stochastically such that

$$\theta^* = \begin{cases} 0, & \text{with probability } \gamma; \text{ and} \\ 0 + \epsilon, & \text{with probability } 1 - \gamma, \end{cases} \tag{16}$$

where:

$$\gamma = \frac{\alpha - \Pr(\theta \leq 0 | data)}{\Pr(\theta < 0 | data) - \Pr(\theta \leq 0 | data)}, \tag{17}$$

and ϵ is an arbitrarily small number.

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:

$$\theta^* = \begin{cases} 0, & \text{with probability } \gamma = 0.959; \text{ and} \\ 0 + \epsilon, & \text{with probability } (1 - \gamma) = 0.041. \end{cases} \tag{18}$$

Defining θ^* in this way will guarantee that, over repeated samples, $\Pr(\theta < \theta^* | data) = 0.05$.

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 could correctly define the credible interval to obtain the correct coverage. When point-null models are involved, model-averaged posteriors are liable to discontinuous point masses and must therefore be approached with 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 both in the incredibility interval, the credible interval must be defined stochastically as either a single point or as an entirely empty interval:

$$(1-\alpha)\%\text{CrI} = \begin{cases} [0], & \text{with probability } \psi; \text{ and} \\ \emptyset, & \text{with probability } (1-\psi), \end{cases} \tag{19}$$

where

$$\psi = \frac{\Pr(\theta = 0|data) - \alpha}{2 \times \Pr(\theta = 0|data) - 1}. \tag{20}$$

When we look at the asymptotic behaviour of these "stochastic credible intervals" it becomes clear that the Jefereys-Lindley paradox predictably reduces the data to be entirely inconsequential. As n increases, γ and ψ both approach $1-\alpha$ since:

$$\begin{aligned} \lim_{n\to\infty}\gamma &= \lim_{n\to\infty}\Big(\frac{\alpha - \Pr(\theta \leq \theta_0|data)}{\Pr(\theta < \theta_0|data) - \Pr(\theta \leq \theta_0|data)}\Big) \\ &= \Big(\frac{\alpha - 1}{-1}\Big) \\ &= 1-\alpha, \end{aligned} \tag{21}$$

and:

$$\begin{aligned} \lim_{n\to\infty}\psi &= \lim_{n\to\infty}\Big(\frac{\Pr(\theta = 0|data) - \alpha}{2 \times \Pr(\theta = 0|data) - 1}\Big) \\ &= \Big(\frac{1-\alpha}{2-1}\Big) \\ &= 1-\alpha. \end{aligned} \tag{22}$$

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 wildly-known consequence of the Jefereys-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 one's credible interval be rather different than the frequentist confidence interval, oftentimes it will be simply undefined (at least in a conventional sense).

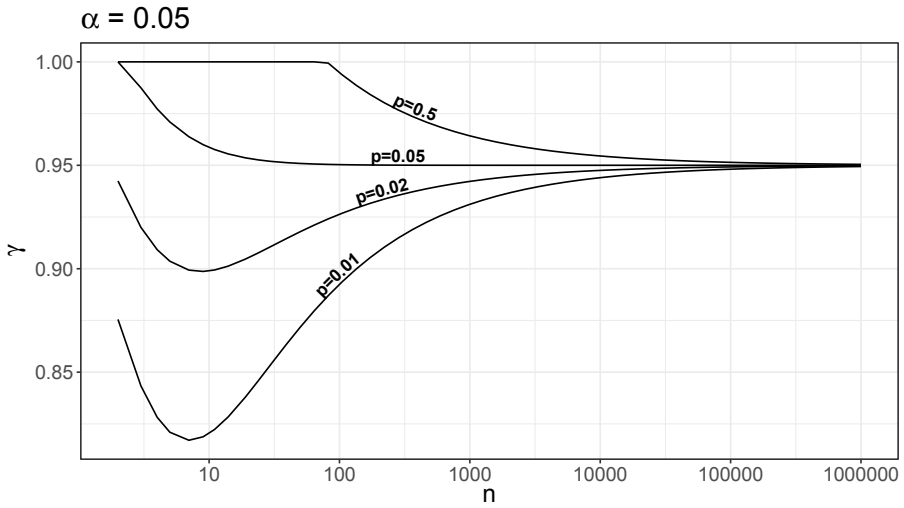

Figure 8: Each line corresponds to observing data corresponding to a p-value of p when testing against $\mathrm{H}_0 : \theta < 0$.

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) for a discussion of Bayesian optimality. However, researchers should be aware that, while perhaps optimal, Bayesian testing with a point-null can lead to rather unexpected asymptotic behaviour.

The consequences of the Jefereys-Lindley paradox on model selection (and null hypothesis testing) are often understood as "intuitive" and not necessarily unfavourable: 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 & Pereira (2016) and Wagenmakers & Ly (2021). However, the consequences of the Jefereys-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.

References

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

Berger, J. O. (1985), *Statistical decision theory and Bayesian analysis*, Springer Science & Business Media.

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

Casella, G. & 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.

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

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. (2022), 'A review of applications of the Bayes factor in psychological research', *Psychological Methods* .

Held, L., Lesaffre, E., Baio, G. & Boulanger, B. (2020), Bayesian tail probabilities for decision making, *in* 'Bayesian Methods in Pharmaceutical Research', CRC Press Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300 . . . , pp. 53–73.

Keysers, C., Gazzola, V. & Wagenmakers, E.-J. (2020), 'Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence', *Nature Neuroscience* **23**(7), 788–799.

Lindley, D. V. (1957), 'A statistical paradox', *Biometrika* **44**(1/2), 187–192.

Pericchi, L. & 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.

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

Wagenmakers, E.-J. & Ly, A. (2021), 'History and nature of the Jeffreys-Lindley paradox', *arXiv preprint arXiv:2111.10191* .

Wasserstein, R. L. & Lazar, N. A. (2016), 'The ASA statement on p-values: context, process, and purpose', *The American Statistician* **70**(2), 129–133.