Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1

Invited Discussion

Bertrand Clarke*

The authors are to be commended for focussing attention on how we understand hypothesis testing – Bayes and frequentist – and its relationship to estimation. These issues seem to have drifted from popular consciousness as we have piled pell-mell onto applied and computational problems. However, now is a good time to revisit these issues and the two examples in this paper provide some much needed clarity.

1 What the Paper Shows

The central point that emerges from this paper is that Bayes factors, credible intervals, and confidence intervals are fundamentally different entities. Sometimes they agree in the sense of leading to equivalent inferences – and sometimes they don't. The interest therefore is mapping out exactly when they agree and understanding why they behave the way they do. Notation here is the same as in the paper.

1.1 Continuous Priors

The example of Sec. 2 shows that Bayes factors and credibility intervals are distinct concepts and that credibility intervals can match confidence intervals with neither matching the inferences from Bayes factors. Consider normal priors $N(0, g_0)$ and $N(0, g_1)$, with $g_0 < g_1$, and a normal likelihood. The Bayes factor in favor of the skinny prior is

$$\begin{aligned} BF_{01} &= \frac{\int \pi_0(\theta) f_{Normal}(\bar{y}, \theta, (1/n)) \mathrm{d}\theta}{\int \pi_1(\theta) f_{Normal}(\bar{y}, \theta, (1/n)) \mathrm{d}\theta} \\ &= \sqrt{\frac{1+ng_1}{1+ng_0}} \exp \frac{(g_0 - g_1) n z^2}{2(1-ng_0)(1-ng_1)}. \end{aligned} \tag{1}$$

Expression (1) can be called a Bayes factor even though it does not correspond to a hypothesis test because it is a fair way to compare two priors; purists might insist on calling it a generalized Bayes factor. Under the asymptotic regime in the paper ($\bar{y} = 1.645/\sqrt{n}$), $Pr(M_0) \to 1/(1 + \sqrt{g_1/g_0})$. For $g_0 = .02$ and $g_1 = 1$ this gives the posterior mixture probability of the model using the skinny prior as .876. That is, the Bayes factor favors the skinny prior relatively strongly. Doubtless, this is so high partially because both priors are normal with a common mean.

Separately, Panel F in Fig. 1 shows that $\Pi(\Theta \geq 0|Data) \geq .5$ and that from Fig. 2 credibility sets asymptotically match confidence sets for θ. The p-value for $\mathcal{H}_0 : \theta < 0$ also asymptotically matches $\Pi(\Theta < 0|data)$ (under the mixture prior). Since this is

arXiv: 2010.00000
*Department of Statistics, University of Nebraska-Lincoln, bclarke3@unl.edu

a continuous case, this is in accord with our intuition. That is, Bayes and frequentist interval estimation matches and the BF is consistent with them.

Overall, the Bayes factor is reflecting the relative appropriateness of the two models i.e., the two priors, but the concept of model selection is disjoint from parameter inference. Indeed, the Bayes factor indicates model sparsity – as we expect from Bayes methods: In the absence of compelling data, the simpler model is preferred.

An interesting feature in Fig. 2 is the hump in the curves around $n = 30$; it is tempting to attribute this to the asymmetry of using $\sqrt{n}\bar{y}$ fixed at a positive value. After all, choosing an asymptotic regime of $z = \sqrt{n}\bar{y}$ with $\bar{y}$ negative might give slightly different curves for lower sample sizes even though they flatten out as n increases.

1.2 Point Mass Priors

The example of Sec. 3 is a limiting case of the example in Sec. 2 where the skinny prior has converged to point mass at zero. This is the setting of the classical Jeffreys-Lindley paradox. That is, as shown in Berger (1980) p. 106-7, when we use a point mass prior we can get statements like $\mathcal{H}_0 : \theta = \theta_0$ is quite believable (from a Bayesian viewpoint) even when the data is five standard deviations away from θ_0.

Specifically, this example shows that if you use a point mass in your prior and carry this over consistently to credible intervals then there is not just a disjunction between frequentist and Bayesian testing, there is also a disjunction between frequentist and Bayesian interval estimation. However, of course, Bayes estimation is in line with the BF as both are based on the posterior. This is the opposite of the continuous case of the example in Sec. 2. For many values of α, the only way to get an interval of the prescribed credibility is to 'split the point' i.e., assign part of the mass at zero to the interval and part to the complement of the interval. This happens because the amount of poterior probability to the right of $\theta = 0$ is too small and the spike at zero is too large for well-defined intervals to exist for arbitrary α. That is, it is not just Bayes testing versus frequentist testing that has a Jeffreys-Lindley paradox, Bayes estimation and frequentist estimation has a Jeffreys-Lindley paradox as well.

More formally, for $\mathcal{H}_0 : \theta = 0$ and $\mathcal{H}_1 : \theta \neq 0$ we get

$$BF_{01} = \sqrt{(1+n)} \exp \frac{-nz^2}{2(1+n)}, \tag{2}$$

which is asymptotically equivalent to $\sqrt{(1+n)}$ when $z = \bar{y}/\sqrt{n}$, meaning we have overwhelming evidence in favor of the null.

In the words of the authors:

> ... due to the discontinuity of the posterior. ... 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.

For comparison purposes, let us also look at two related testing problems. The first is a generalization of the second key point because it involves dimension reduction. Consider two models: Model I has a unidimensional parameter space. Let $\theta \sim N(0, \sigma^2)$ and suppose $(Y_i|\theta) \sim N(\theta, 1)$ are independent for $i = 1, 2$. Model II is two-dimensional: Let $(\theta_1, \theta_2)^T \sim N(0, \sigma^2 I_2)$ and suppose that $(Y_i|\theta) \sim N(\theta_i, 1)$ for $i = 1, 2$. In both cases, we take σ as known. In Model I, the two Y_i's are tied together by a common parameter while in Model II they are not. So rather then reducing a unidimensional parameter space to a zero dimensional parameter space (a point) in Example 2, now we are comparing a two dimensional parameter space with a unidimensional parameter space (the line $\theta_1 = \theta_2$) that is a subset of it.

Let $y = (y_1, y_2)^T$ be the two outcomes and write $\eta = \sigma^2/(2\sigma^2+1)$ and $\tau = \sigma^2/(\sigma^2+1)$. Then, following the derivation in Severinksi et al. (2010), we have

$$\begin{aligned} \frac{m(y|\mathcal{H}_I)}{m(y|\mathcal{H}_{II})} &= \sqrt{2\pi\sigma^2}\frac{\int_{\mathbb{R}} e^{-\frac{1}{2}\sum_{i=1}^2 (y_i-\theta)^2+\theta^2/\sigma^2}}{\Pi_{i=1}^2 \int_{\mathbb{R}} e^{-\frac{1}{2}\sum_{i=1}^2 (y_i-\theta_i)^2+\theta^2/\sigma^2} \mathrm{d}\theta_i} \\ &= \frac{\sigma^2+1}{\sqrt{2\sigma^2+1}} e^{\eta/2(2y_1y_2-\tau(y_1^2+y_2^2))} > 1. \end{aligned} \tag{3}$$

Since (3) is always greater than one we will be led to choose the lower dimensional model. Moreover, if $\sigma \to \infty$, we are led even more strongly to M_I.

That is, like the $N(\theta, g_0)$ vs $N(\theta, g_1)$ example, the lower dimensional model is favored, and we didn't use a true point-null to see this; we used the two-dimensional analog of a point null, a line in the plane. This means that the finding of Example 2 is not likely an anomaly. Moreover, if we remember that shrinkage methods are used to get sparsity and are mathematically equivalent to seeking the mode of a posterior the view that Bayes methods have an inbuilt tendency to sparsity is reinforced.

As another example, closer to Example 1 because the priors are continuous, consider linear models under the Zellner g-prior. Write

$$M_\gamma : Y = \mathbf{1}\beta_0 + X_\gamma \beta_\gamma + \epsilon, \tag{4}$$

for n outcomes, in the usual way, where X_γ is the design matrix with columns corresponding to the γ-th subset of $(X_1, ..., X_p)^T$, $\gamma = (\gamma_1, \ldots, \gamma_p)$ with each γ_j being zero or one indicating the absence or presence of X_j in the model. The vector β contains the regression coefficients with β_γ indicating the entries of β corresponding to the γ_j's equal one. As usual, $\epsilon \sim N(0, \sigma^2 I_p)$. Letting $p_\gamma = \dim(\beta_\gamma)$, assign the Zellner g-prior by choosing $\pi(\beta_\gamma|\gamma) = N_{p_\gamma}(0, g\sigma^2(X_\gamma^T X_\gamma)^{-1})$ and $\pi(\beta_0, \sigma^2|\gamma) = 1/\sigma^2$. Without loss of generality, assume the predictors are centered at zero so that dependence on β_0 is removed. Now, following Clyde and George (2004), cf. Severinksi et al. (2010), we have that the Bayes factor for testing $\mathcal{H}_0$: Null model, $\gamma = \underline{0}$ vs. $\mathcal{H}_1$: Any fixed $\gamma \neq \underline{0}$ is

$$BF(\gamma, 0) = (1+g)^{(n-p_\gamma-1)/2}(1-g(1-R_\gamma^2))^{-(n-1)/2},$$

where R_γ^2 is the usual coefficient of determination.

Suppose the data support γ strongly and we consider the case $R^2_\gamma \to 1$ for fixed values of n and g. Then, $BF(\gamma, 0) \to (1+g)^{(n-p_\gamma-1)/2}$. In the words of Clyde and George (2004):

> ... the Bayes factor ... is bounded no matter how overwhelmingly the data support γ.

Paraphrasing: Even when it is *certain* that the null model is wrong, it still gets nonzero posterior probability, cf. Berger and Pericchi (2001). That is, the Bayes factor always puts some mass on the simplest model. This is much the same as the situation in Example 1 where the Bayes factor allows some weight on both the skinny and fat priors and in the two dimensional example above where the straight line in the joint parameter space is favored.

Moreover, as $g \to 0$ the Zeller g-prior approaches a point mass at zero and it is therefore natural for the null model to be favored by the prior. In this case, the Bayes factor converges to one indicating that the two models are equally favored which makes sense: The data implicit in the prior chooses the null model and the actual data favors γ. In effect, for large n they cancel each other out.

On the other hand, if $g \to \infty$ or $n \to \infty$, i.e., in the limit we are using a flat prior meaning we have essentially no pre-experimental data, the BF increases without bound in response to the actual data and hence we choose γ as we should.

If we were to redo the first example from the paper using a large and small g in the Zellner g-prior, we wouild expect qualitatively identical results and if we were to redo the second example with a convex combination of a point mass prior at, say, zero, and a Zeller g-prior, we would again expect qualitatively identical results. The authors might want to verify this even though it is probably hard to do in closed form.

Incidentally, point nulls or more generally using lower dimensional subspaces of the parameter space seems just fine with the Bayesian approach as long as the dominating measure on the parameter space exists i.e., has a unit mass at a point in it, and densities with respect to it are used consistently.

Problems with point nulls only really arise with frequentist testing because frequentist testing over-rejects; this is a key sense in which Bayes and frequentist testing give fundamentally different results. Consider the familiar χ^2 goodness of fit test for K-cells,

$$\chi^2_{s_1} = \sum_j = 1^K \frac{(O_j - E_j)^2}{E_j} \tag{5}$$

where $E_j = np_j$ and s_1 indicates we have used one sample of size n and O_j is the observed cell count. If we replicate the points exactly to get a data set of size $2n$ (indicated by s_2) and then again to get a data set of size $4n$ (indicated by s_3) the resulting sample χ^2 statistics satisfy

$$\chi^2_{s_3} = 2\chi^2_{s_2} = 2^2\chi^2_{s_1}. \tag{6}$$

Doing this over and over increases the χ^2 statistic and means that any null will eventually be rejected – unless the observed data perfectly matches the expected data.

Frequentists use analogies like comparing the test statistic to a magnifying glass that lets us look ever more closely at the parameter space as the data increase and then eventually saying 'well, yes, we eventually always reject but we didn't really mean a point null to infinite precision anyway'. This is very commonsensical but no way to defend a rigorous theory. The fact is that Bayes and frequentist testing often do not agree and we shouldn't expect them to.

2 Testing: What We Do and Why It's Wrong

Let's be more realistic about how we as Statisticians do our hypothesis testing. The fact is that very, very few of us actually do hypothesis testing properly i.e., follow the procedures that are mathematically justified.

The technical term for this is cheating.

Consider the typical frequentist. Level α testing only makes sense if it's more important to control the probability of Type I errors than the probability of Type II errors, yet most investigators default to this methodology whether justified or not. How many frequentists actually assess the relative costs of the two types of errors? In fact, there are techniques to control both Type I and Type II errors see Cover and Thomas (2006), Chap. 8. So, always using the level α framework is not necesssary.

After this the frequentist finds a p-value; a single number upon which a decision is made. But wait! Frequentists *should* make inferences from the sampling distribution using the notion of confidence. So, to be consistent what they should do is treat the p-value as a random variable and assess its distribution relative to α. Since this is hard, the next best thing would be to bootstrap the p-value. That is, use bootstrapping to generate an empirical distribution for the p-value and then reject only if if the interval formed by p-value $\pm\, 3SE_{p\text{-value}}$ is entirely below (or above) the cutoff level from α for a one-sided test. But, they don't do this.

In the immortal words of Jim Berger: It's one thing to be a frequentist; it's another thing to be a bad frequentist.

And we shouldn't let ourselves off the hook either. All of us know that the optimality of Bayes factors, or equivalently posterior probability, as a criterion for hypothesis testing follows from looking at the posterior risk under generalized zero-one loss, see Berger and Casella (1990) chap. 10. That means the loss of choosing the alternative when the null is true is c_2 and the loss of choosing the null when the alterantive is true is c_1 and the acceptance/rejection threshold for the posterior probability is $c_2/(c_1+c_2)$.

Do we do this? No, we almost never formulate appropriate values for c_1 and c_2 – we set arbitrary thresholds for Bayes factors independent of the problem at hand and de facto assume that the risks associated with the null and alternative are symmetric. Even worse, we generally admit that the generalized zero-one loss is inappropriate since it penalizes deviations equally when usually we think that the loss should be greater the further apart the alternatives are.

Even if our shortcuts are reasonable as a default approach – and many times they

probably are – we should be aware more often of the possibly severe compromises we are making and tell data collectors about it.

3 Being Pragmatic About Testing

If we admit that Bayesian and frequentist statistics are different – per the implications of the examples here showing that outside of smooth priors on equidimensional hypotheses we don't expect agreement – what are we left with?

Quite a lot actually.

Importantly, there really is no Jeffreys-Lindley paradox. It should be called the Jeffreys-Lindley property. Bayesians use a posterior distribution on the parameter space and find probabilities of hypotheses. Frequentists use the sampling distribution on the sample space with the concept of confidence. These techniques are on different measure spaces. Why are we comparing them? Moreover, Bayes and frequentist tests are derived from different optimality criteria. Bayesians use posterior risk whereas frequentists try to maximize density ratios; see the Neyman-Pearson lemma and the most powerful test theory it generates. The surprising bit is that these two procedures ever agree.

Indeed, if there is a paradox, it is that the frequentist optimality criterion for testing looks more reasonable and intuitive than the Bayesian's decision theory criterion but Bayesian procedures work better than frequentist ones. After all, the frequentist looks, fundamentally, at density ratios and puts points in a rejection region that have low density compared to the null density. This is very natural and the centrality of density ratios is accepted by essentially all of us, not just frequentists.

Consider this as a summary of roughly a century of experience with testing: Bayes and frequentist testing are two conceptualy disjoint methodologies that happen to coincide most of the time, at least asymptotically, under strong enough regularity conditions such as parameter spaces that have the same dimension, the use of smooth priors, and conditions to ensure dimension differences between hypotheses don't pose a problem. A *nota bene* to this is that a conditional frequentist testing approach may resolve some of the discrepancies between Bayes and frequentist testing at least partially; see Berger (2003) particularly Sec. 3.3, for a specific procedure. See Fay et al. (2022), especially Sec. 9, for a more recent approach using prior calibration.

Apart from future developments, we can probably also say two things about Bayes versus frequentist testing. First, frequentist testing over-rejects the null, at least relative to Bayes testing. Loosely, a frequentist needs a smaller p-value than expected to justify rejection at least to a Bayesian; see Held and Ott (2018) for more discussion of this.

Second, frequentist methods do function somewhat like a magnification of the parameter space. Bayesian methods do the same but give sparsity as well. For both the frequentist and the Bayesian the 'magnification' rate is $\mathcal{O}(1/\sqrt{n})$ in regular cases so both can distinguish ever smaller differences as n increases. The Bayesian's extra feature of sparsity may be due to the fact that a proper prior ties the whole parameter space together in a way that the samping distribution does not. This may explain why

frequentist testing must be more stringent than Bayes testing in order to achieve the same strength of evidence: behavior in the sampling distribution for hypothesis testing may be a weaker convergence criterion than the posterior provides; see Berger and Delampady (1987).

Thus, Bayes and frequentist testing and estimation are just different and the main reason to hope for a reconciliation is that by so doing we can develop a 'unified field theory for statistics'. Barring future research (and who wants to be pessimistic about that?) the pragmatic approach of using p-values when they are extremely decisive because they are easy to find and otherwise doing a robust Bayesian analysis may be, for now, the best reconciliation available.

References

Berger, J. (1980). *Statistical Decision Theory*. New York: Springer-Verlag. 2

— (2003). "Could Fisher, Jeffreys and Neyman have agreed on testing?" *Stat. Sci.*, 18: 1–32. 6

Berger, J. and Delampady, M. (1987). "Testing precise hypotheses." *Stat. Sci.*, 2: 317–352. 7

Berger, J. and Pericchi, L. (2001). "Objective Bayesian methods for model selection: Introduction and comparison." In *Model Selection*, volume 28 of *Lecture Notes Monograph Series*, 137–193. IMS. 4

Berger, R. and Casella, G. (1990). *Statistical Inference*. Pacific Grove CA: Wadsworth and Brooks/Cole, 1 edition. 5

Clyde, M. and George, E. (2004). "Model uncertainty." *Stat. Sci.*, 19: 81–94. 3, 4

Cover, T. and Thomas, J. (2006). *Elements of Infomation Theory*. Hoboken, NJ: Wiley, 2 edition. 5

Fay, M., Proschan, M., Brittain, E., and Tiwari, R. (2022). "Interpreting p-values and cofidence intervals using well-calibrated null preference priors." *Stat. Sci.*, 37: 455–472. 6

Held, L. and Ott, M. (2018). "On p-values and Bayes factors." *Ann. Rev. Stat. Appl.*, 5: 393–419. 6

Severinksi, C., Fokoué, E., Zhang, H., and Clarke, B. (2010). *Solutions Manual to Principles and Theory for Data Mining and Machine Learning*. New York: Springer. 3