Who Should Be Afraid of the Jeffreys-Lindley Paradox?

Aris Spanos*†

The article revisits the large n (sample size) problem as it relates to the Jeffreys-Lindley paradox to compare the frequentist, Bayesian, and likelihoodist approaches to inference and evidence. It is argued that what is fallacious is to interpret a rejection of H_0 as providing the same evidence for a particular alternative H_1, irrespective of n; this is an example of the fallacy of rejection. Moreover, the Bayesian and likelihoodist approaches are shown to be susceptible to the fallacy of acceptance. The key difference is that in frequentist testing the severity evaluation circumvents both fallacies but no such principled remedy exists for the other approaches.

1. Introduction. The *large n problem* was initially raised by Lindley (1957) in the context of the *simple Normal model*,

$$X_k \sim \text{NIID}(\theta, \sigma^2), \quad k = 1, 2, \ldots, n, \ldots, \tag{1}$$

where $\text{NIID}(\theta, \sigma^2)$ stands for Normal, Independent, and Identically Distributed with mean $\theta \in (-\infty, \infty)$ and variance $\sigma^2 > 0$ (assumed known), by pointing out

> (***a***) **The large *n* problem**. Frequentist testing is susceptible to the fallacious result that there is always a large enough sample size n for which any simple (point) null hypothesis, say H_0: $\theta = \theta_0$, will be rejected by a frequentist α-significance level test.

Received April 2012; revised August 2012.

*To contact the author, please write to: Department of Economics, Virginia Tech, Blacksburg, VA 24061; e-mail: aris@vt.edu.

†Thanks are due to Deborah Mayo for numerous discussions on these topics and to two anonymous referees for many useful comments and suggestions.

Philosophy of Science, 80 (January 2013) pp. 73–93. 0031-8248/2013/8001-0004$10.00

Lindley went on to claim that this result is *paradoxical* because, when viewed from the Bayesian perspective, one can show

> (***b***) **the Jeffreys-Lindley paradox**. For certain choices of the prior, the posterior probability of H_0, given a frequentist α-significance level rejection, will approach one as $n \to \infty$.

This result was later called the *Jeffreys-Lindley paradox* because the broader issue of conflicting evidence between the frequentist and Bayesian approaches was first raised by Jeffreys (1939/1961, 359–60). Claims *a* and *b* contrast the behavior of a frequentist test (p-value) and the posterior probability of H_0 as $n \to \infty$, which brings up a potential for conflict between the frequentist and Bayesian accounts of evidence:

> (***c***) **Bayesian charge 1**. "The Jeffreys-Lindley paradox shows that for inference about θ, p-values and Bayes factors may provide contradictory evidence and hence can lead to opposite decisions" (Ghosh, Delampady, and Samanta 2006, 177).

This potential conflict is given a more distinct Bayesian slant by

> (***d***) **Bayesian charge 2**. A hypothesis that is well supported by the Bayes factor can be (misleadingly) rejected by a frequentist test when n is large (see Berger and Sellke 1987, 112–13; Howson 2002, 45–49).

The problem of conflicting evidence pertains to the broader philosophical issue of grounding statistical practice on sound principles of inference and evidence. What has not been adequately explained in this literature is why, given the rejection of H_0 by a frequentist test, its posterior probability going to one as $n \to \infty$ (irrespective of the truth or falsity of H_0) is conducive to a more sound account of evidence.

The primary objective of this article is to consider this issue by comparing the frequentist, Bayesian, and likelihoodist accounts of inference and evidence. The discussion can be seen as part of a wider endeavor to use the error statistical perspective (Mayo 1996) in an attempt to have a closer look at several Bayesian allegations that have undermined the credibility of frequentist statistics in philosophical circles over the past half century.

The brief comments in section 2 provide a prelude to the discussion that follows by clarifying certain key issues at the outset. Section 3 introduces the large n problem in frequentist testing using a numerical example discussed by Stone (1997). This example is based on a very large sample ($n = 527{,}135$) that is used to bring out the fallacious claims associated with the Jeffreys-Lindley paradox without the technical difficulties of invoking

limiting arguments as $n \to \infty$. In sections 4 and 5, the Bayesian and likelihoodist approaches are applied to this example with a view to demonstrate that both approaches are far from immune to fallacious results, contrary to the current view among proponents of the Bayesian and likelihoodist perspectives (see Berger 1985; Royall 1997; Robert 2007; inter alia). Section 6 illustrates how the postdata severity evaluation of the *p*-value and accept/reject results addresses not only the large *n* problem but also the broader fallacies of acceptance/rejection and calls into question the charges stemming from the Jeffreys-Lindley paradox. The severity perspective is then used in section 7 to shed light on why the Bayesian and likelihoodist accounts of evidence give rise to highly fallacious results.

2. Clarifying What Is Fallacious or Paradoxical. Before discussing the various claims that relate to the Jeffreys-Lindley paradox, it is important to bring out certain key issues that have not been adequately illuminated by the literature.

First, in frequentist testing, which includes both Fisher's significance and the Neyman-Pearson testing, the large *n* problem arises naturally because the power of any "good" (consistent) test increases with *n*. An α-significance level Neyman-Pearson test is said to be consistent when its power to detect any discrepancy $\gamma \neq 0$ from H_0 approaches one as $n \to \infty$. In this sense, there is nothing fallacious or paradoxical about a small *p*-value or a rejection of the null, for a given significance level α, when *n* is large enough, since a highly sensitive test is likely to pick up on tiny (in a substantive sense) discrepancies from H_0.

Second, Bayesian charge 2 (*d*) overlooks the fact that the cornerstone of Neyman-Pearson testing is the trade-off between the type I (reject H_0 when true) and type II (accept H_0 when false) error probabilities. In this sense, these charges ignore the decrease in the type II error probability as *n* increases, since, for a given discrepancy γ, the power is one minus the type II error probability. Indeed, various attempts have been made to alleviate the large *n* problem, including decreasing α as *n* increases in an attempt to counterbalance the increase in power associated with *n* (see Lehmann 1986). The difficulty, however, is that only crude rules of thumb for adjusting α can be devised because the power of a test depends on other factors besides *n*.

Third, the large *n* problem (*a*) constitutes an example of a broader problem known as the *fallacy of rejection*: (mis)interpreting reject H_0 (evidence against H_0) as evidence for a particular H_1; this can easily arise when a test has very high power. Due to the trade-off between type I and II error probabilities, any attempt to ameliorate the problem by selecting a smaller significance level when *n* is large might render the result susceptible to the reverse fallacy known as the *fallacy of acceptance*: (mis)interpreting accept H_0 (no evidence against H_0) as evidence for H_0; this can easily arise when a test

has very low power (e.g., n is very small). As argued below, the large n problem (a), when a rejection of H_0 is interpreted as evidence for H_1, and the Jeffreys-Lindley paradox (b) constitute examples of the fallacies of rejection and acceptance, respectively.

The main argument of this article can be stated succinctly as follows: (i) Although there is nothing fallacious about a small p-value, or a rejection of H_0, when n is large (it is a feature of a good frequentist test), (ii) there is a problem when such results are detached from the test itself and are treated as providing the same evidence for a particular alternative H_1 regardless of the generic capacity (the power) of the test in question. The large n problem is directly related to points (i) and (ii) because the power depends crucially on n. That in turn renders a rejection of H_0 with a large n (high power) very different in evidential terms from a rejection of H_0 with a small n (low power). That is, the real problem does not lie with the p-value or the accept/reject rules as such but with how such results are fashioned into evidence for or against a particular hypothesis or an inferential claim relating to H_0 (H_1). (iii) It is argued that the large n problem can be circumvented by using the postdata severity assessment to provide a sound evidential account for frequentist inference. Whether data $\mathbf{x}_0$ provide evidence for or against a particular hypothesis (H_0 or H_1) depends crucially on the capacity of the test in question to detect discrepancies from the null. This stems from the intuition that a small p-value or a rejection of H_0 based on a test with low power (e.g., a small n) for detecting a particular discrepancy γ provides stronger evidence for the presence of γ than using a test with much higher power (e.g., a large n). As first pointed out by Mayo (1996), this intuition is completely at odds with the Bayesian and likelihoodist intuitions articulated by Berger and Wolpert (1988), Howson and Urbach (2006), and Sober (2008). Indeed, Mayo went on to propose a frequentist evidential account based on harnessing this perceptive intuition in the form of a postdata severity evaluation of the accept/reject results. This is based on custom tailoring the generic capacity of the test to establish the discrepancy γ warranted by data $\mathbf{x}_0$. This evidential account can be used to circumvent the above fallacies, as well as other (misplaced) charges against frequentist testing (see Spanos 2011). (iv) In contrast to frequentist testing, no such reasoned remedy exists for the Bayesian and likelihoodist approaches, whose evidential accounts are shown to be equally vulnerable to these fallacies. A strong case can be made, or so it is argued, that the fallacious nature of the Jeffreys-Lindley paradox stems primarily from the fact that the Bayesian and likelihoodist approaches dismiss the relevance of the generic capacity of the particular test in their evidential accounts. Indeed, these accounts cast aside the relevance of the sample space beyond $\mathbf{x}_0$ and "controlling" the relevant error probabilities for being the key weaknesses of the frequentist approach.

3. The Large *n* Problem in Frequentist Testing. The approach to frequentist statistics followed in this article is known as *error statistics* (Mayo 1996). It can be viewed as a refinement/extension of the Fisher-Neyman-Pearson approach that offers a unifying inductive reasoning for frequentist inference. It refines the Fisher-Neyman-Pearson approach by bringing out the importance of specifying explicitly as well as validating vis-à-vis data $\mathbf{x}_0$ the inductive premises of inference. It extends the Fisher-Neyman-Pearson approach by supplementing it with a postdata severity assessment with a view to address a number of foundational problems relating to a sound evidential account (see Mayo and Spanos 2004, 2006, 2011).

Consider the following numerical example discussed by Stone (1997):

> A particle-physics complex plans to record the outcomes of a large number of independent particle collisions of a particular type, where the outcomes are either type A or type B. . . . The results are to be used to test a theoretical prediction that the proportion of type A outcomes, h, is precisely 1/5, against the vague alternative that h could take any other value. The results arrive: 106,298 type A collisions out of 527,135. (263)

How can one test this substantive hypothesis of interest?

The first step is to embed the above material experiment into a statistical model and frame the substantive hypothesis in terms of statistical parameters. With that in mind, let us assume that each of these n trials can be viewed as a realization of a sample $(X_1, X_2, \ldots, X_n)$ of IID random variables defined by

$$X = \begin{cases} 1 & \text{if type A collision occurs,} \\ 0 & \text{if type B collision occurs,} \end{cases}$$

which transforms the observed sequence of particles into data $\mathbf{x}_0 := (1, 0, 0, 1, 1, \ldots, 0)$. These conditions render the simple Bernoulli (Ber) model,

$$X_k \sim \text{BerIID}[\theta, \theta(1-\theta)], \qquad \theta \in [0,1], \qquad k = 1, 2, \ldots, n, \ldots, \tag{2}$$

appropriate as a statistical model in the context of which the material experiment can be embedded. Note that in practice one should validate the IID assumptions to secure the reliability of inference. The substantive hypothesis of interest, $h = .2$, can be framed in terms of the Neyman-Pearson statistical hypotheses:

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_1: \theta \neq \theta_0, \quad \text{for} \quad \theta_0 = .2, \tag{3}$$

specified solely in terms of the unknown parameter of the statistical model in (2). A proper framing of the Neyman-Pearson hypotheses requires a partitioning of the parameter space, irrespective of whether one is substantively interested in one or more specific values of θ. From the statistical perspective, all values of θ in $[0, 1]$ are relevant for defining the optimality of the test (see Spanos 2010, 569).

The test $T_\alpha := \{d(\mathbf{X}), \mathcal{C}_1(\alpha)\}$, defined by

$$d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{H_0}{\sim} \text{Bin}(0, 1; n),$$
$$\mathcal{C}_1(\alpha) = \{\mathbf{x}: |d(\mathbf{x})| \geq c_{(\alpha/2)}\}, \tag{4}$$

where $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$ and $\mathcal{C}_1(\alpha)$, $c_{(\alpha/2)}$ denote the rejection region and value, respectively, is a uniformly most powerful unbiased Neyman-Pearson test (see Lehmann 1986). Using (4) one can define the type I error probability (significance level):

$$\mathbb{P}(|d(\mathbf{X})| > c_{(\alpha/2)}; H_0) = \alpha. \tag{5}$$

The sampling distribution in (4) is based on the fact that for a Bernoulli IID sample $\mathbf{X} := (X_1, X_2, \ldots, X_n)$, the random variable $Y = n\overline{X}_n = \sum_{i=1}^n X_i$, where Y denotes the number of 1's in n trials, is binomially (Bin) distributed:

$$f(y; \theta, n) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y = 0, 1, 2, \ldots, n.$$

In light of the large sample size ($n = 527{,}135$), it is often judicious to choose a small type I error (Lehmann 1986), say $\alpha = .003$, which yields a rejection value of $c_{(\alpha/2)} = 2.968$; note that the Normal approximation to the binomial distribution is quite accurate in this case.

The power of test T_α at $\theta_1 = \theta_0 + \gamma_1$ defined by

$$\mathcal{P}(\theta_1) = \mathbb{P}(|d(\mathbf{X})| > c_{(\alpha/2)}; \ \theta_1 = \theta_0 + \gamma_1), \quad \text{for } \theta_1 \in \Theta_1,$$

as well as the type II error probability $\beta(\theta_1) = 1 - \mathcal{P}(\theta_1)$, for $\theta_1 \in \Theta_1$, are evaluated using the sampling distribution of $d(\mathbf{X})$ under H_1, which takes the form:

$$d(\mathbf{X}) \overset{\theta=\theta_1}{\sim} \text{Bin}(\delta(\theta_1), V(\theta_1); n), \ \text{for } \theta_1 \in \Theta_1 := \Theta - \{.2\},$$
$$\text{where } \delta(\theta_1) = \frac{\sqrt{n}(\theta_1 - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \quad \text{and} \quad V(\theta_1) = \frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}. \tag{6}$$

This indicates that test T_α is consistent because its power increases with $\delta(\theta_1)$—a monotonically increasing function of $\sqrt{n}$—approaching one as $n \to \infty$ (see Lehmann 1986). Hence, other things being equal, increasing n increases the power of this test, confirming that there is nothing paradoxical about a larger n rendering a (consistent) test more sensitive.

Applying the Neyman-Pearson test T_α to the above data,

$$\begin{aligned} \bar{x}_n &= \frac{106{,}298}{527{,}135} = 0.20165233, \\ d(\mathbf{x}_0) &= \frac{\sqrt{527{,}135}[(106{,}298/527{,}135) - .2]}{\sqrt{.2(1-.2)}} = 2.999, \end{aligned} \tag{7}$$

leads to a rejection of H_0. The p-value, defined as the probability of observing an outcome $\mathbf{x} \in \{0, 1\}^n$ that accords less well with H_0 than $\mathbf{x}_0$ does when H_0 is true, confirms the rejection of H_0:

$$\mathbb{P}(|d(\mathbf{X})| > |d(\mathbf{x}_0)|; H_0) = p(\mathbf{x}_0) = .0027. \tag{8}$$

This definition is preferred to the traditional one, "the probability of observing a result more extreme than $\mathbf{x}_0$ under H_0," because it blends in better with the postdata severity evaluation discussed in section 6. The result $p(\mathbf{x}_0) = .0027$ suggests that data $\mathbf{x}_0$ indicate "some" discrepancy between θ_0 and the "true" θ (that gave rise to $\mathbf{x}_0$), but it provides no information about its magnitude.

As mentioned above, what is problematic is the move from the accept/reject results, and the p-value, to claiming that data $\mathbf{x}_0$ provide evidence for a particular hypothesis, because such a move is highly vulnerable to the fallacies of acceptance and rejection. However, in the context of frequentist testing, this vulnerability can be circumvented using a postdata severity evaluation.

How does the Jeffreys-Lindley paradox arise in this context? Using a *spiked prior* distribution (Lindley 1957) of the form

$$\pi(\theta = \theta_0) = p_0 \quad \text{and} \quad \pi(\theta \neq \theta_0) = 1 - p_0, \tag{9}$$

the formal claim associated with this paradox is

$$\pi\left(H_0 | \overline{X}_n = \theta_0 + c_{(\alpha/2)}\sqrt{\theta_0(1-\theta_0)/n}\right) \to 1 \quad \text{as} \quad n \to \infty;$$

that is, the posterior probability of H_0, conditional on $d(\mathbf{x}_0) = c_{(\alpha/2)}$, goes to one as n approaches infinity. What is not so obvious is why this result is considered an indicator of the soundness of the Bayesian account of evidence.

In relation to the nature of the null hypothesis (simple or composite), it is important to elucidate certain issues raised in this literature. First, in the case of two composite hypotheses

$$H_0\colon \theta \leq \theta_0 \quad \text{versus} \quad H_1\colon \theta > \theta_0, \quad \text{where} \quad \theta_0 = .2, \tag{10}$$

the test $T^*_\alpha := \{d(\mathbf{X}), C^*_1(\alpha)\}$, where $C^*_1(\alpha) = \{\mathbf{x}\colon d(\mathbf{x}) \geq c_\alpha\}$, is both consistent and uniformly most powerful (see Lehmann 1986). In this sense, the large n problem has nothing to do with the restriction to a point null hypothesis; it is simply a feature of a consistent Neyman-Pearson test. The nature of the null (simple or composite) only matters for the behavior of the likelihoodist ratio, the Bayes factor, and the posterior probability of H_0. Second, the fact that the discrepancy between the p-value and the Bayes factor is smaller for a composite null is not particularly interesting because both measures are misleading in different ways.

Third, the comparison of the posterior $\pi(\theta|\mathbf{x}_0)$, as θ varies over $[0, 1]$ (which represent one's revised beliefs about θ in light of $\mathbf{x}_0$), with error probabilities (which measure how often a frequentist procedure errs as $\mathbf{x}$ varies over $\{0, 1\}^n$), is dubious. Indeed, as argued by Casella and Berger (1987, 110), the comparison is made possible by using a blatant misinterpretation of the p-value: "The phrase 'the probability that H_0 is true' has no meaning within frequency theory, but it has been argued that practitioners sometimes attach such a meaning to the p value. Since the p value, in the cases considered, is an upper bound on the infimum of $\Pr(H_0|\mathbf{x})$ it lies within or at the boundary of a range of Bayesian measures of evidence demonstrating the extent to which the Bayesian terminology can be attached."

4. The Bayesian Approach. Consider applying the Bayes factor procedure to the hypotheses (3) using a *uniform prior*:

$$\theta \sim \mathrm{U}(0, 1), \quad \text{that is, } \pi(\theta) = 1 \quad \text{for all } \theta \in [0, 1]. \tag{11}$$

This gives rise to the Bayes factor:

$$\begin{aligned} \mathrm{BF}(\mathbf{x}_0; \theta_0) &= \frac{L(\theta_0; \mathbf{x}_0)}{\int_0^1 L(\theta; \mathbf{x}_0)\, d\theta} \\ &= \frac{\binom{527{,}135}{106{,}298}(.2)^{106{,}298}(1-.2)^{527{,}135-106{,}298}}{\int_0^1 \left(\binom{527{,}135}{106{,}298}\theta^{106{,}298}(1-\theta)^{527{,}135-106{,}298}\right) d\theta} \\ &= \frac{.000015394}{.000001897} = 8.115. \end{aligned} \tag{12}$$

It is interesting to note that the same Bayes factor (12) arises in the case of the spiked prior (9) with $p_0 = .5$, where $\theta = \theta_0$ is given prior probability of .5 and the other half is distributed equally among the remaining values of θ. This is because for $p_0 = .5$, the ratio $(p_0/(1 - p_0)) = 1$ and will cancel out from $\mathrm{BF}(\mathbf{x}_0; \theta)$.

The next step in Bayesian inference is to use (12) as the basis for fashioning an evidential account. A Bayes factor result $\mathrm{BF}(\mathbf{x}_0; \theta_0) > k$, for $k \geq 3.2$, indicates that data $\mathbf{x}_0$ favor the null with the "strength of evidence" increasing with k. In particular, for $3.2 \leq k < 10$, the evidence is substantial, for $10 \leq k < 100$ the evidence is strong, and for $k \geq 100$ is decisive (see Robert 2007).

Comparing the result in (12) with the p-value in (8), Stone (1997, 263) pointed out: "The theoretician is pleased when the [likelihood–Bayes-minded] statistician reports a Bayes factor of 8 to 1 in favour of his brainchild, but the pleasure is alloyed when he uses his own P-value cookbook to reveal the 3.00 standard deviation excess of type A outcomes."

A closer scrutiny of the evidential interpretation of the result in (12) suggests that it is not as clear-cut as it appears. This is because, on the basis of the same data $\mathbf{x}_0$, the Bayes factor $\mathrm{BF}(\mathbf{x}_0; \theta_0)$ "favors" not only $\theta_0 = .2$ but each individual value θ_1 inside a certain interval around $\theta_0 = .2$:

$$\Theta_{\mathrm{BF}} := [.199648,\ .203662] \subset \Theta_1 := \Theta - \{.2\}, \tag{13}$$

where the square bracket indicates inclusion of the end point, in the sense that, for each $\theta_1 \in \Theta_{\mathrm{BF}}$, $\mathrm{BF}(\mathbf{x}_0; \theta_1) > 1$, that is,

$$L(\theta_1; \mathbf{x}_0) > \int_0^1 L(\theta; \mathbf{x}_0) d\theta, \quad \text{for all } \theta_1 \in \Theta_{\mathrm{BF}}. \tag{14}$$

Worse, certain values $\theta^{\ddagger}$ in Θ_{BF} are favored by $\mathrm{BF}(\mathbf{x}_0; \theta^{\ddagger})$ more strongly than $\theta_0 = .2$:

$$\theta^{\ddagger} \in \Theta_{\mathrm{LR}} := (.2, .20331] \subset \Theta_{\mathrm{BF}}, \tag{15}$$

where the left parenthesis indicates exclusion of the end point. It is important to emphasize that the subsets $\Theta_{\mathrm{LR}} \subset \Theta_{\mathrm{BF}} \subset \Theta$ exist for every data $\mathbf{x}_0$, and one can locate them by trial and error. However, there is a much more efficient way to do that. As shown in section 5, Θ_{LR} can be defined as a subset of Θ around the maximum likelihood estimate (MLE) $\widehat{\theta}_{\mathrm{MLE}}(\mathbf{x}_0) = .20165233$. This is not coincidental because, as Mayo (1996, 200) pointed out, $\theta^{\blacklozenge} = \hat{\theta}_{\mathrm{MLE}}(\mathbf{x}_0)$ is always the maximally likely alternative, irrespective of

the null or other substantive values of interest. In this example, the Bayes factor for H_0: $\theta = \theta^{\blacklozenge}$ versus H_1: $\theta \neq \theta^{\blacklozenge}$ yields

$$\text{BF}(\mathbf{x}_0;\theta^{\blacklozenge}) = \frac{\binom{527,135}{106,298}(.20165233)^{106,298}(1-.20165233)^{527,135-106,298}}{\int_0^1 \left(\binom{527,135}{106,298}\theta^{106,298}(1-\theta)^{527,135-106,298}\right)d\theta}$$

$$= \frac{.0013694656}{.000001897} = 721.911, \tag{16}$$

indicating not only decisive evidence for $\theta = \theta^{\blacklozenge}$ but also that

$$\theta^{\blacklozenge} \text{ is favored by BF}(\mathbf{x}_0;\theta^{\blacklozenge}) \text{ more than 89}$$
$$\simeq \frac{721.911}{8.115} \text{ times stronger than } \theta_0 = .2!$$

This result is an instance of the fallacy of acceptance in the sense that the Bayes factor $\text{BF}(\mathbf{x}_0;\theta_0) > 8$ is misinterpreted as providing evidence for H_0: $\theta_0 = .2$ against any value of θ in $\Theta_1 := \Theta - \{.2\}$, when in fact $\text{BF}(\mathbf{x}_0;\theta^{\ddagger})$ provides stronger evidence for certain values of $\theta^{\ddagger}$ in $\Theta_{\text{LR}} \subset \Theta_1$.

What is the source of these conflicting evidence? Stone (1997, 263) conjectured the following explanation: "A subhypothesis strongly rejected by a significance test may be strongly supported in posterior probability if the prior puts insufficient weight on the hypotheses of non-negligible likelihood."

Let us flesh out this conjecture. First, choose $\Theta_{\text{BF}} \subset \Theta_1$ as a range of values of θ that $\text{BF}(\mathbf{x}_0;\theta^{\dagger})$ favors in the sense given in (14). To this range of values the prior attributes insufficient weight since

$$\int_{.199648}^{.203662} d\theta = .004.$$

Second, one can relate Θ_{BF} to both an equal-tail Bayesian $(1-\alpha) = .9997$ credible interval as well as the frequentist $(1-\alpha)$ confidence interval:

$$[\hat{\theta}_{\text{MLE}}(\mathbf{X}) - (3.6267)\text{SD}(\hat{\theta}_{\text{MLE}}(\mathbf{X})),\ \hat{\theta}(\mathbf{X})_{\text{MLE}} + (3.6267)\text{SD}(\hat{\theta}_{\text{MLE}}(\mathbf{X}))],$$

where $\text{SD}(\hat{\theta}_{\text{MLE}}(\mathbf{X}))$ denotes the standard deviation of $\hat{\theta}_{\text{MLE}}(\mathbf{X})$. In light of this, one can consider Θ_{BF} to represent a range of values of θ with nonneg-

ligible likelihood. Third, the weight attributed to these values by the Bayes factor,

$$\int_{.199648}^{.203662} \left(\binom{527{,}135}{106{,}298} \theta^{106{,}298} (1-\theta)^{527{,}135-106{,}298} \right) d\theta = .0000018965, \qquad (17)$$

is rather tiny, providing some support for Stone's conjecture.

This is connected to the large n problem because for $n = 53$ and $y := n\bar{x}_n = 11$ that keeps $\bar{x}_n$ close to its original value, the Bayes factor attribution in (17) would have been much larger ($\gg$) since

$$\int_{.199648}^{.203662} \binom{53}{11} \theta^{11} (1-\theta)^{42} d\theta = .000535 \gg .0000018965.$$

This raises the broader problem of how the large n problem might affect the Bayesian results. In light of the fact that for $n = 53$, $y = 11$,

$$\binom{n}{y} \theta^y (1-\theta)^{n-y} = .13280, \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta = .018518518,$$

the large n problem does not affect the ratio in (12), but it does affect the Bayes attribution by rendering the numerator and denominator much smaller.

Focusing on the latter problem, the question is whether one can address the "insufficient weight" problem, due to the large n, by varying the value of p_0 in $\pi(\theta_0 = .2) = p_0$. That will take an extreme tilting of the prior for a whole range of values of θ to compensate for the particular n. For instance, the tilted spiked-prior

$$\pi(\theta = \theta^\ddagger) = .01 \quad \text{and} \quad \pi(\theta \neq \theta^\ddagger) = .99, \quad \text{for all } \theta^\ddagger \in \Theta_{\text{LR}},$$

can counteract the maximally likely alternative problem. (For further discussion on data-based priors, see Shafer [1982].) Reflecting on this issue, Stone (1997, 264) decried such a Bayesian move, arguing that "if the statistician were to withdraw his uniform prior and claim that he ought to have organized some more probability in the neighbourhood of $\theta = 1/5$, this would be a confession that his Bayesianity does not have a bedrock quality, that his coherence has only the (doubtfully useful) temporal value."

Returning to the invariance of the Bayes factor to the sample size n, it can be shown that it stems from the Fisher-Neyman factorization theorem, where, for a sufficient statistic s for θ, the likelihood function simplifies into

$$L(\theta; \mathbf{x}_0) = f(s; \theta) \times h(\mathbf{x}_0 | s), \quad \text{for all } \theta \in \Theta \qquad (18)$$

(see Cox and Hinkley 1974, 22). In the case of the simple Normal (1) and Bernoulli (2) models, $\overline{X}_n$ is a minimal sufficient statistic for θ, and thus the Bayes factor in (12) depends only on the observed value $\bar{x}_n$. That is, the factor $h(\mathbf{x}_0|s)$ cancels out because it is common to both the numerator and denominator.

Although it seems sensible that the likelihood ratio depends only on $f(s;\theta)$, the claim that it is irrelevant whether $\bar{x}_n$ results from $n = 10$ or $n = 10^{10}$ when going from $\text{BF}(\mathbf{x}_0;\theta_0) > 8$ to claiming that data $\mathbf{x}_0$ provide strong evidence for H_0 seems counterintuitive. As argued in section 7, the large n problem also plagues the Bayes factor primarily because its invariance to n renders its evidential interpretation vulnerable to the fallacy of acceptance, the reverse problem plaguing the p-value. Despite this vulnerability, the likelihood ratio has been proposed as an effective way to deal with the large n problem (see Freeman 1993).

5. The Likelihoodist Approach. The likelihoodist approach (Royall 1997, 24) evaluates how data $\mathbf{x}_0$ compares two simple hypotheses:

$$H_0\colon \theta = \theta_0 \quad \text{versus} \quad H_1\colon \theta = \theta_1,$$

using the likelihood ratio (LR):

$$\text{LR}(\theta_0, \theta_1; \mathbf{x}_0) = \frac{L(\theta_0; \mathbf{x}_0)}{L(\theta_1; \mathbf{x}_0)}. \tag{19}$$

> **Law of Likelihood**: The observations $\mathbf{x}_0$ favor hypothesis H_0 over hypothesis H_1 if and only if $L(\theta_0; \mathbf{x}_0) > L(\theta_1; \mathbf{x}_0)$. And the degree to which $\mathbf{x}_0$ favors H_0 over H_1 is given by the likelihood ratio [19]. (Sober 2008, 32)

It is interesting to note that the Law of Likelihood (LL) was first proposed by Hacking (1965), but a few years later, when reviewing a book on "likelihood," Hacking changed his mind: "The only great thinker who tried it out was Fisher, and he was ambivalent. Allan Birnbaum and myself are very favourably reported in this book for things we have said about likelihood, but Birnbaum has given it up and I have become pretty dubious" (Hacking 1972, 137).

The idea behind the use of the likelihood ratio in (19) is that it ameliorates the large n problem by affecting both hypotheses equally (see Howson and Urbach 2006, 155). Strictly speaking, the LR can only be applied to the case where both hypotheses are simple (see Royall 1997). For hypotheses such as the Neyman-Pearson hypotheses (3), however, the alternative takes an infinite number of values, and thus one needs to select particular point alternatives of interest to apply the LL.

In the case of the above example, a particularly interesting point alternative to $\theta = .2$ is $\theta^{\blacklozenge} = \hat{\theta}_{\text{MLE}}(\mathbf{x}_0)$. For the hypotheses

$$H_0\colon \theta = \theta_0 \quad \text{versus} \quad H_0\colon \theta = \theta^{\blacklozenge},$$

$$\begin{aligned}
&\text{LR}(\theta_0, \theta^{\blacklozenge}; \mathbf{x}_0) \\
&= \frac{\binom{527{,}135}{106{,}298}(.2)^{106{,}298}(1-.2)^{527{,}135-106{,}298}}{\binom{527{,}135}{106{,}298}(.20165233)^{106{,}298}(1-.20165233)^{527{,}135-106{,}298}} \\
&= \frac{.000015394}{.001369466} = .011241,
\end{aligned}$$

which reverses the Bayes factor result and suggests that the degree to which data $\mathbf{x}_0$ favor $\theta = \theta^{\blacklozenge}$ over $\theta_0 = .2$ is much stronger ($89 \simeq (1/.011241)$), confirming the maximally likely alternative problem in (16). In fact, when it comes to fallacious results, the LL is in sync with the Bayes factor procedure because the former can be used directly to establish the subset Θ_{BF} in (13).

To see this, consider pairwise comparisons of different values of $\theta \in \Theta_1$ with $\theta = .2$:

$$H_0\colon \theta = .2 \quad \text{versus} \quad H_1\colon \theta = \theta_1, \quad \text{for all } \theta_1 \in \Theta_1 := \Theta - \{.2\}. \tag{20}$$

The LL reveals that data $\mathbf{x}_0$ favor each value θ_1 in Θ_{LR} over $\theta_0 = .2$ since

$$L(\theta_1; \mathbf{x}_0) > L(\theta_0 = .2; \mathbf{x}_0), \quad \text{for all } \theta_1 \in \Theta_{\text{LR}}.$$

As mentioned above, there is nothing coincidental about the subset $\Theta_{\text{LR}} \subset \Theta_{\text{BF}}$ since

$$\Theta_{\text{LR}} := (.2, .20331] = [\hat{\theta}_{\text{MLE}}(\mathbf{x}_0) \ \pm 3\ \text{SD}(\hat{\theta}_{\text{MLE}})], \tag{21}$$

where

$$\text{SD}(\hat{\theta}_{\text{MLE}}(\mathbf{x}_0)) = \sqrt{\frac{.20165233(1-.20165233)}{527{,}135}} = .0005526.$$

That is, (21) is related to Stone's remark associated with the *p*-value indicating "3 standard deviation excess of type A outcomes," when it is viewed as an observed confidence interval (CI). Having said that, it is important to

reiterate that although we used a CI around $\widehat{\theta}_{\text{MLE}}(\mathbf{x}_0)$ to define Θ_{LR}, such a subset exists in Θ for all data $\mathbf{x}_0$, irrespective of the way one locates Θ_{LR}.

In summary, applying the Bayes factor to the statistical hypotheses,

$$H_0\colon \theta = \theta_0 = .2 \quad \text{versus} \quad H_1\colon \theta = \theta_1, \quad \text{for all } \theta_1 \in \Theta_1 = \Theta - \{.2\}, \tag{22}$$

indicates that data $\mathbf{x}_0$ provides substantial evidence for $\theta_0 = .2$ over $\theta \neq .2$. However, this inference is undermined by the fact that each value $\theta_1 \in \Theta_{\text{LR}} \subset \Theta_1$ turns out to be favored more strongly than $\theta_0 = .2$ when tested using the generic hypotheses:

$$H_0\colon \theta = \theta_0 \quad \text{versus} \quad H_1\colon \theta = \theta_1, \quad \text{for each } \theta_1 \in \Theta_{\text{LR}} \subset \Theta_1, \tag{23}$$

falling prey to the fallacy of acceptance. These conflicting results seriously undermine the initial favoring of $\theta_0 = .2$ over $\theta \neq .2$, and they bring out the fallacious implications of Bayesian and likelihoodist inferences, calling into question the Jeffreys-Lindley paradox. As aptly put by Stone (1997, 263): "When not misused, they [*P*-values] still provide some sort of control over the pursuit of weak clues—not a measure of faith in some alternative hypothesis. A *P*-value is a *P*-value is a *P*-value! That some users like to misinterpret it as a posterior probability or odds ratio or other inferential measure . . . should not detract from the *P*-value's intrinsic, if uninterpretable, value."

As shown in the next section, in the case of the *p*-value, there is a principled way to circumvent its weaknesses using Mayo's (1996) postdata severity assessment.

6. Severity: Addressing the Large *n* Problem. Severity constitutes a postdata evaluation of the Neyman-Pearson accept/reject results with a view to establish the (smallest/largest) discrepancy γ from H_0 warranted by data $\mathbf{x}_0$. As such, the severity evaluation is by definition directional since postdata one has an outcome $d(\mathbf{x}_0)$ whose sign indicates the relevant direction of departure from H_0.

A hypothesis H passes a severe test T_α with data $\mathbf{x}_0$ if (i) $\mathbf{x}_0$ accords with H (using a suitable measure of accordance), and (ii) with very high probability, test T_α would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false.

The case of testing the hypotheses (3) using (4) yielded $d(\mathbf{x}_0) = 2.999$, which led to rejecting H_0. Postdata, the sign of the observed test statistic, $d(\mathbf{x}_0) > 0$, indicates that the rejection is clearly in the direction of values greater than $\theta_0 = .2$. That is, in light of data $\mathbf{x}_0$, the two-sidedness of the original Neyman-Pearson test is irrelevant. Indeed, the same severity evalu-

ation applies to the case of the one-sided test for the composite hypotheses (10). Condition (i) of severity implies that the generic form of the inferential claim that "passed" is (Mayo and Spanos 2006):

$$\theta > \theta_1 = \theta_0 + \gamma, \quad \text{for some } \gamma \geq 0. \tag{24}$$

It is important to emphasize that (24) is not a reformulation of the original hypotheses but a framing of the relevant inferential claim associated with the rejection of H_0 stemming from $d(\mathbf{x}_0) = 2.999$ (see Spanos 2011). The rejection of H_0 calls for the appraisal of the largest discrepancy γ from H_0 warranted by data $\mathbf{x}_0$. Condition (ii) calls for the evaluation of the probability of "outcomes that accord less well with $\theta > \theta_1$ than $\mathbf{x}_0$ does," which translates into all those outcomes $\mathbf{x} \in \{0, 1\}^n$ such that $[d(\mathbf{x}) \leq d(\mathbf{x}_0)]$. This gives rise to:

$$\text{SEV}(T_\alpha; \theta > \theta_1) = \mathbb{P}(\mathbf{x}: d(\mathbf{X}) \leq d(\mathbf{x}_0); \theta > \theta_1 \text{ is false}). \tag{25}$$

Given the numerical values in (7) and the relevant distribution in (6), one can proceed to evaluate (25) for different $\theta_1 > \theta_0$ using the standardized statistic:

$$\frac{d(\mathbf{X}) - \delta(\theta_1)}{\sqrt{V(\theta_1)}} \overset{\theta=\theta_1}{\sim} \text{Bin}(0, 1; n), \quad \text{for } \theta_1 > \theta_0. \tag{26}$$

Table 1 reports the severity evaluation for different discrepancies of interest.

To explain how one derives the results in table 1, consider the case $\gamma = .002$, that is, the relevant claim is $\theta > .202$. Evaluating the relevant components

$$\frac{d(\mathbf{x}_0) - \delta(\theta_1)}{\sqrt{V(\theta_1)}} = \frac{2.999 - 3.63}{\sqrt{1.0007}} = -.631,$$

$$\delta(\theta_1) = \frac{\sqrt{527135}(.202 - .2)}{\sqrt{.2(1 - .2)}} = 3.63,$$

$$V(\theta_1) = \frac{.202(1 - .202)}{.2(1 - .2)} = 1.0007,$$

one can proceed to evaluate $\text{SEV}(T_\alpha; \theta > \theta_1)$ using the $N(0, 1)$ tables, which yields:

$$\text{SEV}(T_\alpha; \theta > \theta_1) = \mathbb{P}(\mathbf{x}: d(\mathbf{X}) \leq -.631; \theta = .202) = .264.$$

TABLE 1. REJECT H_o: $\theta = (d(xo) = 2.999)$

γ	Inferential Claim $\theta > \theta_1 = \theta_0 + \gamma$,	SEV$(T_\alpha;\ \theta > \theta_1)$ $\mathbb{P}(\mathbf{x}: d(\mathbf{X}) \leq -d(X_0);\ \theta = \theta_1)$
0	$\theta > .2000$	.999
.0009	$\theta > .2009$	.914
.00095	$\theta > .20095$	.900
.001	$\theta > .2010$	.882
.0015	$\theta > .2015$	.609
.00165	$\theta > .20165$	.500
.002	$\theta > .2020$	.264
.0023	$\theta > .2023$	.120
.0024	$\theta > .2024$	.087
.0025	$\theta > .2025$	.062
.003	$\theta > .203$	.007

The results in table 1 indicate that, for a severity threshold of say .9, the claim for which data $\mathbf{x}_0$ provide evidence is $\theta > .20095 \Rightarrow \gamma^* \leq .00095$.

How does this answer relate to the original question of interest of testing the theoretical prediction that the proportion of type A outcomes is .2? One needs to answer the question whether the particular discrepancy from the null, γ^*, is substantively significant, which cannot be answered exclusively on statistical grounds because it pertains to the substantive subject matter information. That is, one needs to consider γ^* in the context of the theory of particle physics that motivated the above experiment to decide whether it is substantively significant.

It is important to emphasize that the postdata severity evaluation goes beyond avoiding the misuse of *p*-values, as suggested by Stone (1997) in the above quotation. It addresses the key problem with Fisherian *p*-values in the sense that the severity evaluation provides the "magnitude" of the warranted discrepancy from the null by taking into account the generic capacity of the test (that includes n) in question as it relates to the observed data $\mathbf{x}_0$.

As shown in Mayo and Spanos (2006), the postdata severity assessment can be used to supplement frequentist testing with a view to bridge the gap between the *p*-value and the accept/reject rules, on the one hand, and providing evidence for or against a hypothesis in the form of the discrepancy γ from the null warranted by data $\mathbf{x}_0$, on the other hand. Its key difference from the Bayesian and likelihoodist approaches is that it takes into account the generic capacity of the test in establishing γ.

The severity-based evidential interpretation addresses not just the large-n problem but the fallacies of acceptance and rejection more broadly, as well as other charges leveled against frequentist testing, including the "arbitrariness"

of choosing the significance level, the one-sided versus two-sided framing of hypotheses, the reversing of the null and alternative hypotheses, and so forth (see Mayo and Spanos 2011; Spanos 2011).

7. Bayesian and Likelihoodist Accounts Revisited. Where does the above severity assessment leave the Bayesian and likelihoodist inferences? Both approaches are plagued by the *maximally likely alternative* problem (Mayo 1996) in the sense that the value $\widehat{\theta}_{\text{MLE}}(\mathbf{x}_0) = \theta^{\blacklozenge}$ is always favored against every other value of θ, irrespective of the substantive values of interest. Any attempt to sidestep that problem will require an extreme data-based tilting of the prior against all values $\theta^{\ddagger} \in \Theta_{\text{LR}}$. In contrast, the severity of the inferential claim $\theta > \theta^{\blacklozenge}$ is always low, being equal to .5 (table 1), calling into question the appropriateness of such a choice. In addition, the severity assessment in table 1 calls seriously into question the results associated with the two intervals $\Theta_{\text{BF}} := [.199653, .203662]$ and $\Theta_{\text{LR}} := (.2, .20331]$ because these intervals include values $\theta^{\ddagger}$ of θ, for which the severity of the relevant inferential claim $\theta > \theta^{\ddagger}$ is very low (e.g., $\text{SEV}(T_{\alpha}; \theta > .2033) \simeq .001$).

The question that naturally arises is why the Bayesian and likelihoodist approaches give rise to the above conflicting and confusing results. The severity account gives a straightforward answer: both approaches ignore the generic capacity of a test when going from

$$\textbf{step 1}\text{: } \text{LR}(\theta_0, \theta_1; \mathbf{x}_0) = \frac{L(\theta_0; \mathbf{x}_0)}{L(\theta_1; \mathbf{x}_0)} > k, \tag{27}$$

indicating that θ_0 is k times more likely than θ_1, to **step 2:** fashioning the result in (27) into the strength of evidence for or against θ_i, $i = 0, 1$.

Bayesians and likelihoodists are likely to challenge this criticism as misplaced by invoking the distinction between the logic versus the epistemology of inference to claim that the generic capacity of the test belongs to the latter, but their approaches are primarily focused on the former. Demarcating the logic of inference as pertaining to what follows from given premises (in a deductive sense) and the epistemology of inference as concerned with "how we learn from data $\mathbf{x}_0$," the generic capacity of a test belongs squarely within the logic of inference because it follows deductively from the premises (the statistical model) without any reference to data $\mathbf{x}_0$. The inductive dimension of severity emanates from the fact that its evidential account uses the particular data $\mathbf{x}_0$ to infer something pertaining to the underlying generating mechanism represented by the statistical model.

The Bayesian objection would have had some merit if the approach were to end after $\text{BF}(\mathbf{x}_0; \theta_0)$ is evaluated, but it does not. Similarly, likelihoodists do not end after $\text{LR}(\theta_0, \theta_1; \mathbf{x}_0)$ is evaluated but proceed to claim an evidential

based on benchmarks for the "strength of statistical evidence" for θ_0. For instance, $\text{LR}(\theta_0, \theta_1; \mathbf{x}_0) > k$, $k = 8$, is considered moderate evidence, while $k = 32$ is considered strong evidence (see Royall 1997). In contrast, the severity account ensures learning from data $\mathbf{x}_0$ by employing reliable procedures to establish trustworthy evidence for hypotheses or claims pertaining to the underlying generating mechanism, the reliability of inference being calibrated in terms of the relevant error probabilities, both predata and postdata.

What is particularly interesting is that the Bayes factor and the likelihood ratio are directly related to the test (4) in the sense that the test statistic $d(\mathbf{X})$ is a monotone function of the likelihood ratio statistic $\lambda(\theta; \mathbf{X}) = [L(\theta_0; \mathbf{X}) / \max_{\theta \in \Theta} L(\theta; \mathbf{X})]$, and its rejection region $C_1(\alpha)$ is related to $\lambda(\theta; \mathbf{X}) > k$ (see Lehmann 1986). In this sense, the key difference between the frequentist and the other two approaches is that they ignore the sampling distribution and the associated error probabilities of the test in (4) by invoking the likelihood principle (Berger and Wolpert 1988), which asserts that no other value $\mathbf{x}$ of the sample $\mathbf{X}$, apart from data $\mathbf{x}_0$, is relevant for inference purposes. Indeed, Bayesian statisticians take delight in poking fun at frequentist testing by quoting Jeffreys's (1939/1961, 385) remark about the "absurdity" of invoking the quantifier "for all $\mathbf{x} \in \{0, 1\}^n$": "What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure."

What these critics overlook is that their attempts to provide an evidential account for statistical hypotheses go astray exactly because they ignore the generic capacity of the test, which calls upon the quantifier "for all $\mathbf{x} \in \{0, 1\}^n$" for its evaluation. Viewed from the severity perspective, the trouble with using a small p-value as a basis for inferring evidence for a particular alternative H_1 stems from the fact that it only indicates the presence of "some" discrepancy from H_0, but it provides no information about its magnitude; the latter requires summoning the generic capacity of the test. In light of this fatal flaw of the p-value as a basis for an evidential account, the literature concerned with "reconciling" the p-value (or some modification of it) with various Bayesian measures (see Berger and Delampady 1987; Berger, Boukai, and Wang 1997; Sellke, Bayarri, and Berger 2001; Berger 2003; inter alia) is overlooking the real issue. Any evidential account aiming to provide a sound answer to the question of when data $\mathbf{x}_0$ provide evidence for or against a hypothesis (or a claim) can ignore the generic capacity of a test at its peril!

Intuitively, what goes wrong is that the Bayesian factor and the likelihoodist procedures use Euclidean geometry based on $\text{LR}(\theta_0, \theta_1; \mathbf{x}_0)$ to evaluate evidence for different hypotheses (H_0 or H_1), when in fact the statistical testing space is curved, with the curvature determined primarily by the capac-

ity of the test to detect discrepancies from H_0. This is especially relevant for the large n problem because the power of the test in (4) increases with $\sqrt{n}$. The use of Euclidean geometry is clearly apparent in the following quotation from Berger and Wolpert (1988, 7):

> For a testing example, suppose it is desired to test
>
> $$H_0\colon \mu = -1 \text{ versus } H_1\colon \mu = 1, \qquad [28]$$
>
> based on $X \sim N(\mu, .25)$. The rejection region $X \geq 0$ gives a test with error probabilities (type I and type II) of .0228. If $x = 0$ is observed, it is then permissible to state that H_0 is rejected and that the error probability is $\alpha = .0228$. Common sense, however, indicates that $x = 0$ fails to discriminate at all between H_0 and H_1. Any sensible person would be equally uncertain as to the truth of H_0 or H_1 (based just on data $x = 0$).

Any "sensible" Bayesian/likelihoodist would be wrong to conclude that $x = 0$ provides the same evidence for both H_0 and H_1 just because it is halfway between their hypothesized values of μ (see Spanos 2011).

Although there is a connection between the "curved statistical space" and Stone's (1997) conjecture concerning "insufficient weight on the hypotheses of non-negligible likelihood," there is no way to use the generic capacity of the test to provide a more appropriate weighting scheme within the Bayesian and likelihoodist frameworks.

8. Summary and Conclusions. The Jeffreys-Lindley paradox has played an important role in undermining the credibility of frequentist inference by focusing attention on how the large-n problem renders frequentist testing vulnerable to the fallacy of rejection. It was argued that although there is nothing fallacious about rejecting H_0, when n is large, there is a problem when this result is detached from the test itself and viewed as providing the same evidence for a particular alternative H_1 regardless of the generic capacity (that depends on n) of the test in question. This renders the p-value and the accept/reject rules vulnerable to the fallacies of acceptance and rejection.

The discussion has also called into question the basic premise of the Jeffreys-Lindley paradox concerning the sagacity of the Bayes factor favoring H_0 as n increases as symptomatic of the fallacy of acceptance, the reverse problem plaguing the p-value. More generally, it was shown that the move from $[L(\theta_0; \mathbf{x}_0)/L(\theta_1; \mathbf{x}_0)] > k$ to inferring that $\mathbf{x}_0$ provides weak or strong evidence for H_0, depending on the value of $k > 1$, renders the Bayes factor and the likelihood ratio equally susceptible to the same fallacies.

It was argued that in the context of frequentist testing these fallacies can be circumvented using a postdata severity evaluation. The key is that this evaluation takes into account the test's generic capacity in establishing the

discrepancy γ from the null warranted by data $\mathbf{x}_0$. The underlying intuition is that detecting a particular discrepancy γ using a very sensitive (insensitive) test provides less (more) strong evidence that γ is present. In contrast, the Bayesian and likelihoodist approaches have no principled way to circumvent these fallacies.

REFERENCES

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer.

———. 2003. "Could Fisher, Jeffreys, and Neyman Have Agreed on Testing?" *Statistical Science* 18 (1): 1–32.

Berger, James O., Ben Boukai, and Yinping Wang. 1997. "Unified Frequentist and Bayesian Testing of Precise Hypotheses." *Statistical Science* 12 (3): 133–60.

Berger, James O., and Mohan Delampady. 1987. "Testing Precise Hypotheses." *Statistical Science* 2 (3): 317–35.

Berger, James O., and Thomas Sellke. 1987. "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence." *Journal of the American Statistical Association* 82 (397): 112–22.

Berger, James O., and Robert W. Wolpert. 1988. *The Likelihood Principle*. Lecture Notes, Monograph Series, 2nd ed, vol. 6. Hayward, CA: Institute of Mathematical Statistics.

Casella, George, and Roger L. Berger. 1987. "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem." *Journal of the American Statistical Association* 82 (397): 106–11.

Cox, David R., and David V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman & Hall.

Freeman, P. R. 1993. "The Role of *P*-Values in Analyzing Trial Results." *Statistics in Medicine* 12:1433–59.

Ghosh, Jayanta K., Mohan Delampady, and Tapas Samanta. 2006. *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer.

Hacking, Ian. 1965. *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

———. 1972. Review of *Likelihood: An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference*, by A. F. Edwards. *British Journal for the Philosophy of Science* 23:132–37.

Howson, Colin. 2002. "Bayesianism in Statistics." In *Bayes's Theorem*, ed. R. Swinburne, 39–71. Oxford: Oxford University Press.

Howson, Colin, and Peter Urbach. 2006. *Scientific Reasoning: A Bayesian Approach*. 3rd ed. Chicago: Open Court.

Jeffreys, Harold. 1939/1961. *Theory of Probability*. Oxford: Oxford University Press.

Lehmann, Erich L. 1986. *Testing Statistical Hypotheses*. 2nd ed. New York: Wiley.

Lindley, Dennis V. 1957. "A Statistical Paradox." *Biometrika* 44:187–92.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G., and Aris Spanos. 2004. "Methodology in Practice: Statistical Misspecification Testing." *Philosophy of Science* 71:1007–25.

———. 2006. "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction." *British Journal for the Philosophy of Science* 57:323–57.

———. 2011. "Error Statistics" In *Philosophy of Statistics, Handbook of Philosophy of Science*, ed. D. Gabbay, P. Thagard, and J. Woods, 153–98. Burlington: Elsevier.

Robert, Christian. 2007. *The Bayesian Choice*. 2nd ed. New York: Springer.

Royall, Richard M. 1997. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Sellke, Thomas, M. J. Bayarri, and James O. Berger. 2001. "Calibration of *P*-Values for Testing Precise Null Hypotheses." *American Statistician* 55:62–71.

Shafer, Glenn. 1982. "Lindley's Paradox." *Journal of the American Statistical Association* 77:325–34.

Sober, Elliott. 2008. *Evidence and Evolution: The Logic behind the Science*. Cambridge: Cambridge University Press.
Spanos, Aris. 2010. "Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?" *Philosophy of Science* 77:565–83.
———. 2011. "Misplaced Criticisms of Neyman-Pearson (N-P) Testing in the Case of Two Simple Hypotheses." *Advances and Applications in Statistical Science* 6:229–42.
Stone, Mervyn. 1997. "Discussion of Aitkin (1997)." *Statistics and Computing* 7:263–64.