PHIL6334/ECON6614-Lecture Notes 5: Revisiting the Jeffreys-Lindley Paradox

Aris Spanos [Spring 2019]

1 Introduction

The apparent paradox was initially raised by Lindley (1957) in the context of the simple Normal model:

$$\mathcal{M}_\theta(\mathbf{x}): \quad X_k \backsim \mathsf{NIID}(\theta, \sigma^2), \; k{=}1,2,...,n,... \tag{1}$$

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

The problem raised by Lindley pertains to the *p-value* as it compares to the *Bayes factor* inference as they relate to testing the N-P hypotheses:

$$H_0\text{: } \theta{=}\theta_0 \text{ vs. } H_1\text{: } \theta{\neq}\theta_0,$$

where the **p-value** takes the form:

$$\mathbb{P}(|d(\mathbf{X})|>|d(\mathbf{x}_0)|\,;\; \theta{=}\theta_0), \tag{2}$$

where $d(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma}$, $\overline{X}_n{=}\frac{1}{n}\sum_{i=1}^n X_i$, and the **ratio of posteriors** is:

$$\tfrac{\pi(H_0|\mathbf{x}_0)}{\pi(H_1|\mathbf{x}_0)}{=}\tfrac{L(H_0|\mathbf{x}_0)}{L(H_1|\mathbf{x}_0)}\left(\tfrac{\pi(H_0)}{\pi(H_1)}\right),$$

where $B_{01}{=}\frac{L(H_0|\mathbf{x}_0)}{L(H_1|\mathbf{x}_0)}$ denotes the Bayes factor and $(\pi(H_0)/\pi(H_1))$ the ratio of the priors.

Lindley (1957) pointed out an *old issue* raised initially by Berkson (1938):

[**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 for any $0{<}\alpha{<}1$,

but 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 priors $\pi(H_0)$ and $\pi(H_1)$, such as:

$$\pi(\theta{=}\theta_0){=}p_0 \text{ and } \pi(\theta{\neq}\theta_0){=}1-p_0, \; 0<p_0<1, \tag{3}$$

known as a *spiked prior,* the posterior probability $\pi(H_0|\mathbf{x}_0)$, given a frequentist α-significance level rejection, i.e. $\pi(H_0 \mid d(\mathbf{x}_0){=}c_{\frac{\alpha}{2}})$, will approach one as $n{\to}\infty$, i.e.

$$\pi(H_0 \mid \overline{x}_n{=}\theta_0+c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})) \underset{n\to\infty}{\longrightarrow} 1 \tag{4}$$

NOTE that the implicit rejection region for the N-P test is:

$$C_1(\alpha) = \{\mathbf{x}\!:\ |d(\mathbf{x})| \geq c_{\frac{\alpha}{2}}\},$$

and H_0 is rejected at the boundary $d(\mathbf{x}_0)=\frac{\sqrt{n}(\overline{x}_n-\theta_0)}{\sigma}=c_{\frac{\alpha}{2}} \rightarrow \overline{x}_n=\theta_0 + c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$.

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), pp. 359-360.

The claims [a] and [b], pertaining to the behavior as $n\rightarrow\infty$ of a frequentist test (p-value) and the posterior probability of H_0, respectively, stem from two different perspectives on testing, that creates a potential for conflict between their respective 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 et. al, 2006, p. 177)

This potential conflict is given a more distinct Bayesian slant by the charge:

[d] **Bayesian charge 2**: a hypothesis that is well-supported by Bayes factor can be (misleadingly) rejected by a frequentist test when n is large; see Berger and Sellke (1987), pp. 112-3, Howson (2002), pp. 45-9.

The problem of conflicting evidence pertains to the broader philosophical issue of grounding statistical practice on sound principles of inference and evidence.

The primary objective of this paper is to consider this question 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) to revisit several Bayesian allegations that have undermined the credibility of frequentist statistics in philosophical circles over the last half century.

2 Clarifying what is paradoxical and what is not!

Before we discuss 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 this literature.

First, in frequentist testing, which includes both Fisher's significance and the Neyman-Pearson (N-P) testing, the large n problem arises naturally because the power of any 'good' (consistent) test increases with n. An α-significance level N-P test is said to be *consistent* when its power to detect any discrepancy $\gamma \neq 0$ from H_0 approaches one as $n \rightarrow \infty$. 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, the Bayesian charge in [d] overlooks the fact that the cornerstone of N-P 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, like 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 issue 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 (see [a]), when a rejection of H_0 is interpreted as evidence *for* H_1, and *the Jeffreys-Lindley paradox* (see [b]), constitute examples of the fallacies of rejection and acceptance, respectively.

Fourth, are the two probabilities:

$$\mathbb{P}(|d(\mathbf{X})| > |d(\mathbf{x}_0)|\,;\ \theta{=}\theta_0){=}p(\mathbf{x}_0)\ \text{ vs. }\ \pi(\theta{=}\theta_0 \mid \overline{x}_n{=}\theta_0 + c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})),$$

even comparable in any way that can be considered legitimate from a frequentist perspective? The p-value $p(\mathbf{x}_0)$ refers to the two tail areas of the sampling distribution of $d(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sigma}$, when $\theta{=}\theta_0$ is hypothetically assumed to be true. It evaluates the smallest significance level α at which H_0 would have been rejected when H_0 is true, with the probablities firmly attached to $\mathbf{X}$ as it varies in the sample space $\mathbb{R}^n$. The posterior probability $\pi(\theta{=}\theta_0 |\ \overline{x}_n)$ denotes the revised prior probability $\pi(\theta{=}\theta_0)$ in light of the particular value $\overline{x}_n{=}\theta_0 + c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$.

As argued by Casella and Berger (1987), p. 110, the comparison is made possible by using a blatant misinterpretation of the p-value:

"The phrase "the probability that H_0 is true" [$\pi(\theta{=}\theta_0)$] has no meaning within frequency theory."

Fifth, the **real paradox** is why, given the rejection of H_0 by an α-level frequentist test, its posterior probability goes to one as $n{\to}\infty$, *irrespective of the truth or falsity* of H_0, is conducive to a more sound account of evidence? It is not! This can be explained from a frequentist perspective by pointing out that the result

$$\pi(H_0 \mid \overline{x}_n{=}\theta_0 + c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})) \underset{n\to\infty}{\longrightarrow} 1$$

is *trivially true* since $c_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}} \underset{n\to\infty}{\longrightarrow} 0$, by invoking the SLLN:

$$\mathbb{P}(\lim_{n\to\infty}(\overline{X}_n{-}\theta_0 - c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}){=}0){=}1,\ \text{ i.e. }\ \overline{X}_n \xrightarrow[n\to\infty]{a.s.} \theta_0.$$

But that implicitly assumes that $\theta_0=\theta^*$, since the SLLN asserts that:

$$\mathbb{P}(\lim_{n\to\infty}\overline{X}_n=\theta^*)=1,$$

when the IID assumptions are valid for data $\mathbf{x}_0$. That is, Lindley's argument stems from assuming that $\theta_0=\theta^*$, irrespective of whether it **is true or false** for data $\mathbf{x}_0$. Caution: it is one thing to use hypothetical reasoning based $\theta=\theta_0$ to evaluate how often a N-P test errs, and completely another to assume that factually $\theta_0=\theta^*$.

The *main argument* that follows can be stated succinctly summarized:

(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 context (the statistical model, the test, the hypotheses formulation, etc.), 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.

(iii) This problem can be adequately addressed using the post-data severity assessment to provide an evidential account for frequentist inference.

(iv) No such reasoned remedy exists for the Bayesian and likelihoodist approaches whose evidential accounts are shown to be are equally vulnerable to these fallacies.

The large n problem is directly related to these claims 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).

Whether data $\mathbf{x}_0$ provide evidence for or against a particular hypothesis H (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 intuition 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 **post-data 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.

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.

3 The large n problem in frequentist testing

The approach to frequentist statistics followed in this paper is known as *error statistics*; see Mayo (1996). It can be viewed a refinement/extension of the Fisher-Neyman-Pearson (F-N-P) approach that offers a unifying inductive reasoning for frequentist inference. It extends the F-N-P approach by supplementing it with a *post-data severity* assessment with a view to address a number of foundational problems bedeviling frequentist inference since the 1940s; see Mayo and Spanos (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: 106298 type A collisions out of 527135." (p. 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 $(X_1, X_2, ..., X_n)$ 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** $\mathbf{x}_0{:=}(1, 0, 0, 1, 1, ..., 0)$. These conditions render the simple Bernoulli (Ber) model:

$$X_k \backsim \mathsf{BerIID}(\theta, \theta(1{-}\theta)),\ \theta{\in}[0, 1],\ k{=}1, 2, ..., n, ... \tag{5}$$

appropriate as a statistical model in the context of which the material experiment can be embedded. The substantive hypothesis of interest, $h{=}.2$, can be framed in terms of the Neyman-Pearson (N-P) statistical hypotheses:

$$H_0\text{: } \theta = \theta_0 \quad \text{vs.} \quad H_1\text{: } \theta \neq \theta_0, \text{ for } \theta_0{=}.2, \tag{6}$$

specified solely in terms of the unknown parameter(s) of the statistical model (5). As argued in Spanos (2010), p. 569, the proper framing of the N-P 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.

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

$$d(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{H_0}{\backsim} \mathsf{Bin}\left(0, 1; n\right), \qquad \mathcal{C}_1(\alpha){=}\{\mathbf{x}\text{: } |d(\mathbf{x})| \geq c_{\frac{\alpha}{2}}\}, \tag{7}$$

where $\overline{X}_n{=}\frac{1}{n}\sum_{i=1}^{n} X_i$ and $\mathcal{C}_1(\alpha)$, $c_{\frac{\alpha}{2}}$ denote the rejection region and value, respectively, is a Uniformly Most Powerful Unbiased N-P test (see Lehmann, 1986). Using (7) one can define the type I error probability (significance level):

$$\mathbb{P}(|d(\mathbf{X})| > c_{\frac{\alpha}{2}}; H_0) = \alpha. \tag{8}$$

The sampling distribution in (7) is based on the fact that for a Bernoulli IID sample $\mathbf{X}{:=}(X_1, X_2, ..., X_n)$, 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:

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

In light of the large sample size (n=527135), it is often judicious to choose a smaller type I error (Lehmann, 1986), say α=.003, which yields a rejection value of $c_{\frac{\alpha}{2}}$=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:

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

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:

$$\begin{gathered} d(\mathbf{X}) \overset{\theta=\theta_1}{\backsim} \mathsf{Bin}(\delta(\theta_1), V(\theta_1); n), \text{ for } \theta_1{\in}\Theta_1{:=}\Theta - \{.2\}, \\ \text{where } \delta(\theta_1){=}\tfrac{\sqrt{n}(\theta_1-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \text{ and } V(\theta_1){=}\tfrac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}. \end{gathered} \tag{9}$$

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 (good) test more sensitive.

Applying the N-P test T_α to the above data:

$$\overline{x}_n{=}\tfrac{106298}{527135}{=}0.20165233, \quad d(\mathbf{x}_0){=}\frac{\sqrt{527135}(\frac{106298}{527135}-.2)}{\sqrt{.2(1-.2)}}{=}2.999, \tag{10}$$

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:

$$\mathbb{P}(|d(\mathbf{X})| > |d(\mathbf{x}_0)|; H_0) = p(\mathbf{x}_0){=}.0027. \tag{11}$$

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 accords better with the post-data 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 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 H, 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 post-data severity evaluation; see section 6.

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

$$\pi(\theta=\theta_0)=p_0 \text{ and } \pi(\theta\neq\theta_0)=1-p_0, \tag{12}$$

the formal claim associated with this paradox is:

$$\pi\left(H_0|d(\mathbf{x}_0)=c_{\frac{\alpha}{2}}\right)\rightarrow 1 \text{ as } n\rightarrow\infty, \tag{13}$$

i.e. the posterior probability of H_0, conditional on or $\overline{x}_n=\theta_0+c_{\frac{\alpha}{2}}\sqrt{\theta_0(1-\theta_0)/n}$, goes to one as n approaches infinity. What is *not* so obvious is 'why is this result considered a virtue for the Bayesian account of evidence?' As argued above, (13) implicitly assumes the validity of the null, i.e. $\theta_0=\theta^*$, which is an unwarranted claim!

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.

4 The Bayesian approach

Consider applying the *Bayes factor* procedure to the hypotheses (6) using a *uniform prior*:

$$\theta\backsim\mathsf{U}(0,1), \text{ i.e. } \pi(\theta)=1 \text{ for all } \theta\in[0,1]. \tag{14}$$

This gives rise to the Bayes factor:

$$BF(\mathbf{x}_0;\theta_0)=\frac{L(\theta_0;\mathbf{x}_0)}{\int_0^1 L(\theta;\mathbf{x}_0)d\theta}=\frac{\binom{527135}{106298}(.2)^{106298}(1-.2)^{527135-106298}}{\int_0^1\left(\binom{527135}{106298}\theta^{106298}(1-\theta)^{527135-106298}\right)d\theta}=\frac{.000015394}{.000001897}=8.115. \tag{15}$$

It is interesting to note that the same Bayes factor (15) arises in the case of the *spiked prior* (12) with p_0=.5, where $\theta=\theta_0$ is given prior probability of .5 and 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 $BF(\mathbf{x}_0;\theta)$.

The next step in Bayesian inference is to use (15) as the basis for fashioning an evidential account. A Bayes factor result $BF(\mathbf{x}_0;\theta_0)>k$, for $k\geq 3.2$, indicates that data $\mathbf{x}_0$ *favors* 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 (15) with the p-value in (11), Stone (1997) 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." (p. 263)

A closer scrutiny of the evidential interpretation of the result in (15) suggests that it is not as clear cut as it appears. This is because, on the basis of same data $\mathbf{x}_0$, the

Bayes factor $BF(\mathbf{x}_0;\theta_0)$ 'favors', not only θ_0=.2, but each individual value θ_1 inside a certain interval around θ_0=.2:

$$\Theta_{BF}{:=}[.199648,\ .203662]\subset\Theta_1{:=}\Theta-\{.2\},\tag{16}$$

where the square bracket indicates inclusion of the end point, in the sense that:

$$\text{for each } \theta_1{\in}\Theta_{BF},\ BF(\mathbf{x}_0;\theta_1){>}1, \text{ i.e. }\ \ L(\theta_1;\mathbf{x}_0)>\int_0^1 L(\theta;\mathbf{x}_0)d\theta,\ \text{ for all } \theta_1{\in}\Theta_{BF}\tag{17}$$

Worse, certain values $\theta^{\ddagger}$ in Θ_{BF} are favored by $BF(\mathbf{x}_0;\theta^{\ddagger})$ *more strongly* than θ_0=.2:

$$\theta^{\ddagger}{\in}\Theta_{LR}{:=}(.2,\ .20331]\subset\Theta_{BF},\tag{18}$$

where the curly bracket indicates exclusion of the end point. It is important to emphasize that the subsets $\Theta_{LR}\subset\Theta_{BF}\subset\Theta$ 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) $\widehat{\theta}_{MLE}(\mathbf{x}_0)$=.20165233. This is not coincidental because as Mayo (1996), p. 200, pointed out, $\theta^{\blacklozenge}{=}\widehat{\theta}_{MLE}(\mathbf{x}_0)$ 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: $\theta{=}\theta^{\blacklozenge}$ vs. H_1: $\theta{\neq}\theta^{\blacklozenge}$ yields:

$$BF(\mathbf{x}_0;\theta^{\blacklozenge})\ \ =\frac{\binom{527135}{106298}(.20165233)^{106298}(1-.20165233)^{527135-106298}}{\int_0^1\left(\binom{527135}{106298}\theta^{106298}(1-\theta)^{527135-106298}\right)d\theta}\ \ =\frac{.0013694656}{.000001897}{=}721.911,\tag{19}$$

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

$\theta^{\blacklozenge}$ is favored by $BF(\mathbf{x}_0;\theta^{\blacklozenge})$ more than $89{\simeq}\frac{721.911}{8.115}$ times stronger than θ_0=.2!

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

What is the source of these conflicting evidence? Stone (1997) conjectured the following explanation:

> "a subhypothesis is 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." (p. 263)

Let us flesh out this conjecture in three steps. *First*, choose $\Theta_{BF}\subset\Theta_1$ as a range of values of θ which $BF(\mathbf{x}_0;\theta^{\dagger})$ favors in the sense given in (17). 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:

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

where $SD(\widehat{\theta}_{MLE}(\mathbf{X}))$ denotes the Standard Deviation (SD) of $\widehat{\theta}_{MLE}(\mathbf{X})$. In light of this, one can consider Θ_{BF} to represent a range of values of θ with 'non-negligible likelihood'. *Third*, the weight attributed to these values by the Bayes factor:

$$\int_{.199648}^{.203662}\left(\binom{527135}{106298}\theta^{106298}(1-\theta)^{527135-106298}\right)d\theta=.0000018965, \tag{20}$$

is rather tiny, providing some support for Stone's conjecture. This is connected to the large n problem, because for n=53 and y:=$n\overline{x}_n$=11 that keeps $\overline{x}_n$ close to its original value, the Bayes factor attribution in (20) would have been much larger since:

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

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:

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

the large n problem does *not* effect the ratio in (15), 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:

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

can counteract the *maximally likely alternative* problem. For further discussion on data-based priors see Shafer (1982). Reflecting on this issue, Stone (1997), p. 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:

$$L(\theta;\mathbf{x}_0)=f(s;\theta)\cdot h(\mathbf{x}_0|s),\ \text{for all } \theta\in\Theta; \tag{21}$$

see Cox and Hinkley (1974), p. 22. In the case of the simple Normal (1) and Bernoulli (5) models, $\overline{X}_n$ is a minimal sufficient statistic for θ, and thus the Bayes factor in (15) depends only on the observed value $\overline{x}_n$; 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 $\overline{x}_n$ results from n=10 or n=10^{10} when going from $BF(\mathbf{x}_0;\theta_0)>8$ 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.

5 The likelihoodist approach

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

$$H_0\text{: } \theta{=}\theta_0 \text{ vs. } H_1\text{: } \theta{=}\theta_1,$$

using the *Likelihood Ratio (LR)*: $LR(\theta_0, \theta_1; \mathbf{x}_0) = \frac{L(\theta_0;\mathbf{x}_0)}{L(\theta_1;\mathbf{x}_0)}$. (22)

"**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 (22)." (Sober, 2008, p. 32).

It is interesting to note that the Law of Likelihood (LL) was first proposed by Hacking (1965), but later on 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, p. 137)

The idea behind the use the Likelihood Ratio (LR) in (22) is that it ameliorates the large n problem by affecting both hypotheses equally; see Howson and Urbach (2006), p. 155. Strictly speaking, the LR can only be applied in the case where both hypotheses are simple. For hypotheses such as (6), however, the alternative takes an infinite number of values, and thus one needs to select particular point alternatives of interest to apply the Law of Likelihood (LL).

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

$$H_0\text{: } \theta{=}\theta_0 \text{ vs. } H_0\text{: } \theta{=}\theta^{\blacklozenge},$$

$$LR(\theta_0, \theta^{\blacklozenge}; \mathbf{x}_0){=}\frac{\binom{527135}{106298}(.2)^{106298}(1{-}.2)^{527135-106298}}{\binom{527135}{106298}(.20165233)^{106298}(1{-}.20165233)^{527135-106298}}{=}\frac{.000015394}{.001369466}{=}.011241,$$

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 ($89{\simeq}\frac{1}{.011241}$), confirming the *maximally likely alternative* problem in (19). In fact, when it comes to fallacious results, the LL is in total agreement with the Bayes factor procedure because the former can be used directly to establish the subset Θ_{BF} in (16).

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

$$H_0\text{: } \theta{=}.2 \text{ vs. } H_1\text{: } \theta{=}\theta_1, \text{ for all } \theta_1{\in}\Theta_1{:=}\Theta - \{.2\}. \tag{23}$$

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

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

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

$$\Theta_{LR}{:=}(.2,\ .20331] = [\widehat{\theta}_{MLE}(\mathbf{x}_0) \pm 3SD(\widehat{\theta}_{MLE})], \tag{24}$$

where $SD(\widehat{\theta}_{MLE}(\mathbf{x}_0))=\sqrt{\frac{.20165233(1-.20165233)}{527135}}=.0005526$. (24) is related to Stone's remark associated with the p-value indicating '3 standard deviation excess of type A outcomes', when (24) is viewed as an observed Confidence Interval (CI). Having said that, it is important to re-iterate that although we used a CI around $\widehat{\theta}_{MLE}(\mathbf{x}_0)$ to define Θ_{LR}, such a subset exists in Θ for every data $\mathbf{x}_0$, irrespective of the way one uses to locate it.

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

$$H_0\text{: } \theta = \theta_0 = .2 \quad \text{vs.} \quad H_1\text{: } \theta{=}\theta_1, \text{ for all } \theta_1{\in}\Theta_1{:=}\Theta - \{.2\}, \tag{25}$$

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 $\theta_1{\in}\Theta_{LR}{\subset}\Theta_1$ turns out to be favored more strongly than $\theta_0{=}.2$ when tested using the generic hypotheses:

$$H_0\text{: } \theta = \theta_0 \quad \text{vs.} \quad H_1\text{: } \theta = \theta_1, \text{ for each } \theta_1{\in}\Theta_{LR} \subset \Theta_1, \tag{26}$$

falling prey to the *fallacy of acceptance*. These conflicting results seriously undermine the initial favoring of $\theta_0{=}.2$ over $\theta{\neq}.2$, and bring out the fallacious implications of Bayesian and likelihoodist inferences, calling into question the Jeffreys-Lindley paradox. As aptly put by Stone (1997):

> "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." (p. 263)

Indeed, in the case of the p-value there is a principled way to circumvent its weaknesses using Mayo's (1996) post-data severity assessment.

6 Severity: addressing the large n problem

Severity constitutes a post-data evaluation of the N-P accept/reject results with a view to establish the (smallest/largest) discrepancy γ from the null hypothesis warranted by data $\mathbf{x}_0$. As such, the severity evaluation is by definition *directional* since post-data one has an outcome $d(\mathbf{x}_0)$ indicating the 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),
(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.

In the case of testing the hypotheses (6) using (7) yielded $d(\mathbf{x}_0){=}2.999$, which led to *rejecting* H_0. Post-data, 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, post-data (in light of data $\mathbf{x}_0$) the two-sidedness of the original N-P test is irrelevant. Indeed, the same severity evaluation applies to the case of the one sided test for the composite hypotheses (**??**). Condition (i) of severity implies that the generic form of the inferential claim that 'passed' is (Mayo and Spanos, 2006):

$$\theta > \theta_1 = \theta_0 + \gamma, \quad \text{for some } \gamma \geq 0. \tag{27}$$

It is important to emphasize that (27) 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>0; 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 the event '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:

$$SEV(T_\alpha;\theta > \theta_1)=\mathbb{P}(\mathbf{x} : d(\mathbf{X}) \leq d(\mathbf{x}_0); \theta > \theta_1 \text{ is false}). \tag{28}$$

Given the numerical values in (10) and the relevant distribution in (9), one can proceed to evaluate (28) for different $\theta_1 > \theta_0$ using the standardized statistic:

$$\tfrac{d(\mathbf{X})-\delta(\theta_1)}{\sqrt{V(\theta_1)}} \overset{\theta=\theta_1}{\backsim} \mathsf{Bin}(0,1;n), \text{ for } \theta_1 > \theta_0. \tag{29}$$

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

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

$$\delta(\theta_1)=\tfrac{\sqrt{527135}(.202-.2)}{\sqrt{.2(1-.2)}}=3.63, \quad V(\theta_1)=\tfrac{.202(1-.202)}{.2(1-.2)}=1.0007, \quad \tfrac{d(\mathbf{x}_0)-\delta(\theta_1)}{\sqrt{V(\theta_1)}}=\tfrac{2.999-3.63}{\sqrt{1.0007}} = \text{-.631},$$

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

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

The results in table 1 indicate that, for a severity threshold of say .9, the claim for which data $\mathbf{x}_0$ provide evidence *for* is:

$$\theta > \ .20095 \Rightarrow \gamma^* \leq \ .00095.$$

Table 1: Reject H_0: θ=.2 ($d(\mathbf{x}_0)$=2.999)

	Inferential claim	$SEV(T_\alpha;\theta > \theta_1)$
γ	$\theta > \theta_1$=θ_0+γ,	$\mathbb{P}(\mathbf{x}$:$d(\mathbf{X}) \leq d(\mathbf{x}_0); \theta$=$\theta_1)$
0	$\theta > .2000$	.999
.0008	$\theta > .2008$	.939
.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

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 which motivated the above experiment to decide whether it is substantively significant or not.

It is important to emphasize that the post-data 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; 2011) the post-data 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 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. 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 underlying intuition is that detecting a discrepancy γ using a very sensitive (insensitive) test provides less (more) strong evidence that γ is present; see Mayo (1996).

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 vs. two-sided formulations of hypotheses, the reversing of the null and alternative hypotheses, etc.; see 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 $\widehat{\theta}_{MLE}(\mathbf{x}_0)=\theta^{\blacklozenge}$ 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 $\theta^{\ddagger}\in\Theta_{LR}$. 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 Θ_{BF}:=[.199653, .203662] and Θ_{LR}:=(.2, .20331], because they include values $\theta^{\ddagger}$ of θ for which the severity of the relevant claim $\theta > \theta^{\ddagger}$ is very low, e.g. $SEV(T_\alpha; \theta > .2033) \simeq .001$.

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: $$\textit{step 1:}\quad LR(\theta_0,\theta_1;\mathbf{x}_0)=\tfrac{L(\theta_0;\mathbf{x}_0)}{L(\theta_1;\mathbf{x}_0)}>k, \tag{30}$$

indicating that θ_0 is k times more likely than θ_1, to *step 2*: fashioning the result in (30) into the strength of evidence *for* or *against* $\theta_i,\ i{=}0,1$.

Bayesians and likelihoodists are likely to challenge this criticism as misplaced by invoking the distinction between the logic vs. 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 a 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 stems from the fact that its evidential account uses the particular data $\mathbf{x}_0$ to infer something pertaining to the 'true' generating mechanism represented by the statistical model.

The Bayesian objection would have had some merit if the approach were to end after $BF(\mathbf{x}_0;\theta_0)$ is evaluated, but it doesn't. Similarly, likelihoodists do not end after $LR(\theta_0,\theta_1;\mathbf{x}_0)$ is evaluated, but proceed to claim an evidential account (epistemology) based on benchmarks for the 'strength of statistical evidence' for θ_0. For instance, $LR(\theta_0,\theta_1;\mathbf{x}_0){>}k$, $k{=}8$ 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 pre-data and post-data.

What is particularly interesting is that the Bayes factor and the likelihood ratio are directly related to the test (7) in the sense that the test statistic $d(\mathbf{X})$ is a monotone function of the likelihood ratio *statistic* $\lambda(\theta;\mathbf{X}){=}\frac{L(\theta_0;\mathbf{X})}{\max_{\theta\in\Theta}L(\theta;\mathbf{X})}$, and its rejection region $\mathcal{C}_1(\alpha)$ 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 (7), 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) 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." (p. 385)

What these critics overlook is that their attempts to provide an *evidential account* for statistical hypotheses goes 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 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, Sellke, et. al, 2001, Berger, 2003, inter alia) is overlooking the real issue:

any *evidential account* aiming to provide a sound answer the question: 'when do 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!

8 Summary and conclusions

Although there is nothing fallacious about rejecting H_0, when n is large, there *is* a problem when this result is **detached** from its context, and viewed as providing the same evidence for a particular alternative H_1, regardless of the generic capacity (the power) of the test in question. Such a practice 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.

A case is made that the **real paradox** is raised by the Bayesian account of evidence since the posterior probability of H_0 given $d(\mathbf{x}_0)=c_{\frac{\alpha}{2}}$:

$$\pi(H_0 \mid \overline{x}_n=\theta_0+c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})) \underset{n\rightarrow\infty}{\longrightarrow} 1 \tag{31}$$

irrespective of the truth or falsity of H_0. As argued above (31) is *trivially true* since $c_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}} \underset{n\rightarrow\infty}{\longrightarrow} 0$, by invoking the SLLN that asserts:

$$\mathbb{P}(\lim_{n\rightarrow\infty}\overline{X}_n=\theta^*)=1,$$

when the IID assumptions are valid for data $\mathbf{x}_0$. That is, Lindley's argument stems from assuming that $\theta_0=\theta^*$, irrespective of whether it **is true or false** for data $\mathbf{x}_0$.

It was argued that in the context of frequentist testing these fallacies can be easily addressed using a post-data severity evaluation; see Mayo and Spanos (2006). 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 $\mathbf{x}_0$. In contrast, the Bayesian and likelihoodist approaches have no principled way to circumvent these fallacies.