# On the Jeffreys–Lindley's paradox*

CHRISTIAN P. ROBERT

Université Paris-Dauphine, CEREMADE, University of Warwick, Department of Statistics, and CREST, Paris

*Abstract.* This paper discusses the dual interpretation of the Jeffreys–Lindley's paradox associated with Bayesian posterior probabilities and Bayes factors, both as a differentiation between frequentist and Bayesian statistics and as a pointer to the difficulty of using improper priors while testing. We stress the considerable impact of this paradox on the foundations of both classical and Bayesian statistics. While assessing existing resolutions of the paradox, it focus on a critical viewpoint of the paradox discussed by Spanos (2013) in the current journal.

*Key words and phrases:* Bayesian inference, Testing statistical hypotheses, Type I error, significance level, p-value.

## 1. INTRODUCTION

In the statistical literature, there is little debate as to whether or not testing statistical hypotheses is the most controversial aspect of statistical inference, with at least three major competing schools approaching the problem from different angles and often concluding with opposite decisions. In this regard, Lindley's (1957) paradox may constitute the most quoted instance of the opposition between the frequentist and Bayesian schools of inference. Two recent reassessments of the paradox appeared in *Philosophy of Science*, with Spanos (2013) and Sprenger (2013) diverging in their resolution of the paradox, which prompted me to reconsider in turn this fundamental argument both in the frequentist-Bayesian debate and in the derivation of (more) coherent testing procedures within the Bayesian framework.

Let me first recall the setting of the paradox as exposed in Lindley (1957), often called *the Jeffreys–Lindley's paradox* after Dennis Lindley pointed out the facts were already exposed in Jeffreys (1939). Given a sample of size $n$ from a normal distribution $\mathcal{N}(\theta, \sigma^2)$ with known variance $\sigma^2$, testing whether or not the null hypothesis $H_0 : \theta = \theta_0$ on the mean holds (against the alternative $H_1 : \theta \neq \theta_0$) may lead to opposite conclusions depending on the statistical perspective adopted

*Christian P. Robert, CEREMADE, Université Paris-Dauphine, 75775 Paris cedex 16, France xian@ceremade.dauphine.fr and Department of Statistics, University of Warwick, Coventry CV4 8UW, UNited Kingdom. Research partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2012–2015 grant ANR-11-BS01-0010 "Calibration" and by an Institut Universitaire de France chair. The author is most sincerely grateful to the editorial team for its comments and support towards improving the exposition in the paper.

to conduct the test. Namely, summarising the dataset into the sufficient statistic

$$\bar{x}_n \sim \mathcal{N}(\theta, \sigma^2/n)$$

leads to the $t$ statistic $t_n = \sqrt{n}(\bar{x}_n - \theta_0)/\sigma$ which is distributed as a $\mathcal{N}(0,1)$ variable under the null hypothesis, allowing for the computation of the $p$-value equal to

$$p(t_n) = \mathbb{P}(|T_n| > |t_n|) = 1 - 2\Phi(|t_n|)\,.$$

Relying upon this $p$-value to determine whether or not $H_0$ holds means examining its value either in absolute terms or with respect to a bound. A Bayesian approach to the hypothesis testing problem, as exposed in Jeffreys (1939) relies on the ratio of evidences (or marginal likelihoods) also called the Bayes factor (see Berger, 1985 or Robert, 2001). When the prior distribution on the parameter $\theta$ is the normal prior, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, the Bayes factor is given by

$$\mathfrak{B}_{01}(t_n) = (1+n)^{1/2} \exp\left(-nt_n^2/2[1+n]\right)\,, \tag{1}$$

which measures the evidence brought by the data in favour of the null hypothesis relative to the alternative hypothesis. A decision about which hypothesis to select is then based on the numerical value of $\mathfrak{B}_{01}(t_n)$, the default boundary between null and alternative being $\mathfrak{B}_{01}(t_n) = 1$.

The paradox exposed by Lindley (1957) is that, *for a fixed numerical value of $t_n$ and for almost any choice of prior distribution on $\theta$, the Bayes factor $\mathfrak{B}_{01}(t_n)$ goes to infinity with the sample size $n$ while the $p$-value $p(t_n)$ remains constant in $n$.* In Lindley's words, "we [can be] 95% confident [as frequentists] that $\theta \neq \theta_0$ but have 95% belief [as Bayesians] that $\theta = \theta_0$" (p.187). This occurs for instance when $t_n = 1.96$ and $n = 16,818$ assuming the prior weights equally both hypotheses. (And for $n = 164$ if $H_0$ is ten times more likely than $H_1$.) Sprenger (2013) takes the example of a test for extra-sensory capacities (ESP) to oppose a $p$-value of 0.003 and a Bayes factor of 12 in favour of the null. This divergence of outcomes is called a "paradox" since the same dataset almost certainly leads to opposite conclusions and hence decisions when the sample size $n$ is large. It led many commentators of the paradox to conclude that one approach or the other was "wrong".

While divergences between different statistical theories of inference and their numerical conclusions are to be expected, the fact that they persist when the sample size grows to infinity explains for the long-term impact of this paradox and the fact that it is still the focus of attention for statisticians and philosophers of science alike. Although the frequentist-Bayesian opposition expressed by the paradox can be thoroughly explained, as detailed in Section 2.1, the Jeffreys-Lindley's paradox also has consequences within the Bayesian framework that are detailed in this paper.

My personal understanding of the Jeffreys–Lindley's paradox is that it points at the poor and even unacceptable behaviour of improper prior distributions when testing point-null hypotheses. See, e.g., the resolution proposed in Robert, 1993, aimed at suppressing the impact of an arbitrary normalising constant in improper priors. However, a large majority of quotes and comments view the paradox as an irreconcilable divergence between the Bayesian and the frequentist resolutions of the point-null hypothesis testing problem, blaming (at least) one of those approaches for the discrepancy. I will thus start the paper with an analysis of the opposition between the $p$-value and the Bayes factor (see, e.g., Kass and Wasserman, 1996). I want to stress at this early stage that my Bayesian approach follows the decision-theoretic perspective of Berger (1985), which means that hypothesis testing is conducted with the intent of a course of action (depending on the selection or rejection of $H_0$) rather than for the epistemic attempt of uncovering the "truth", agreeing in this respect with the position advocated in Sprenger (2013).

The plan of this paper is as follows: I review the points for and against the paradox in Section 2, analyse the recent criticism of Spanos (2013) in Section 3, discuss some Bayesian resolutions of the paradox in Section 4, and conclude in Section 5.

## 2. SETTING THE PARADOX REFERENTIALS

### 2.1 Frequentist versus Bayesian interpretations

Let us recall that the classical view of the Jeffreys–Lindley's paradox is that the Bayes factor and the $p$-value asymptotically (in the sample size) differ to the point of leading to opposite conclusions (acceptance versus rejection of the null hypothesis $H_0$).

There is obviously no mathematical issue with the paradox—otherwise it would have been readily dismissed—: as the quantities involved in the two perspectives evaluate different objects using different measures: the probability measure of an event over the sample space versus the probability measure of an event over the parameter space, the former being conditional on the parameter value and the later on the observation of the sample. Despite the large literature on the topic, I would also argue that this is *not* a statistical paradox in that observing a constant value of[1] $t_n$ as $n$ increases is not of statistical interest: when $H_0$ is true, $t_n$ has a limiting $\mathcal{N}(0,1)$ distribution, which means the corresponding $p$-value has a limiting uniform distribution, while, when $H_0$ does not hold, $t_n$ converges almost surely to $\infty$, in which case both the Bayes factor and the $p$-value converge to 0. This behaviour is completely in line with the general result of the consistency of the Bayes factor in this setting, which is all too often overlooked in most commentaries on the Jeffreys–Lindley's paradox. And the Neyman–Pearson (frequentist) approach to testing suggests decreasing both the Type I and Type II error, hence the acceptance boundary for the $p$-value when $n$ increases (see, e.g., Lehmann and Casella, 1988).

There are several reasons why the two approaches, Bayesian _and frequentist ($\mathfrak{f}$), should not numerically agree, even asymptotically:

(a) one _operates on the parameter space $\Theta$, range of the possible values of $\theta$ under the alternative, while the other ($\mathfrak{f}$) is produced exclusively on the sample space $\mathcal{X}$; this is the same opposition between confidence ($\mathfrak{f}$) and credibility _when constructing interval estimators (a point also made by Sprenger, 2013);

(b) one ($\mathfrak{f}$) relies solely on the point-null hypothesis $H_0$ and the sampling distribution it induces, while the other _opposes the null $H_0$ to a marginal

[1] As pointed out by Lindley (1957): "5% in to-day's small sample does not mean the same as 5% in to-morrow's large one" (p.189).

version of $H_1$ (in that it is integrated over the parameter space $\Theta$ against a specific prior distribution);

(c) following what may be the most famous quote from Jeffreys (1939, Section 7.2) one ($\mathfrak{f}$) could rejects "a hypothesis that may be true (...) because it has not predicted observable results that have not occurred" ($\{X > x_{\text{obs}}\}$, say), while the other _conditions upon the observed value $x_{\text{obs}}$;

(d) at least in the Fisherian perspective, one ($\mathfrak{f}$) resorts to an arbitrary fixed bound $\alpha$ on the $p$-value, while the other _refers to the boundary probability of $1/2$ (unless a genuine loss function or an unbalanced prior weighting vector is constructed).

A consequent literature (see, e.g. Berger and Sellke, 1987) has since then shown how divergent those two approaches could prove (to the point of being asymptotically incompatible). Despite the fact that both approaches are consistent in the sense mentioned above, most commentators on the paradox conclude to the worthlessness of the $p$-value (always rejecting at a given $\alpha$ level for $n$ large enough), or of the Bayes factor (always accepting for a fixed $p$-value for $n$ large enough), or of both (see, e.g., Spanos, 2013, and Sprenger, 2013). Others (see, e.g., Gelman et al. (2013)) have chosen to bypass the opposition by considering tools at the interface between both approaches, like posterior predictive checks.

## 2.2 Improper inputs for Bayes factors

While the gap between the frequentist and the Bayesian degrees of evidence was the reason for Lindley (1957) mentioning a statistical paradox, an orthogonal consequence of Jeffreys's (1939) and Lindley's (1957) exhibiting this paradox is to highlight the genuine difficulty in using improper priors[2] in testing settings: as stressed by Lindley (1957), "the only assumption that will be questioned is the assignment of a prior distribution of any type" (p.188). This was also the argument made by both Shafer (1982) and DeGroot (1982) (see also DeGroot, 1973) in their discussion of the paradox. Note that Jeffreys does not address the general problem of using improper priors in testing, using ad-hoc solutions when available and developing a second (and under-appreciated) type of Jeffreys's priors otherwise (see Robert et al., 2009, Section 6.4, for a discussion).

This second (but not secondary) level of interpretation for the paradox shifts the asymptotics from the sample size to a prior scale factor. If we remain within the normal framework of Lindley (1957), with one observation $x \sim \mathcal{N}(\theta, \sigma^2)$, considering a prior distribution of the form $\theta \sim \mathcal{N}(\theta_0, n\sigma^2)$ under the alternative hypothesis leads to a Bayes factor that is identical to (1) for $t_n = x$. In this perspective, $n$ is a prior scale factor, so that the prior variance is $n$ times larger than the observation variance.[3]. The interpretation of the phenomenon is then obviously different: *when the prior scale $n$ goes to infinity, the Bayes factor goes to infinity no matter what the value of the observation $x$ is*. (Note that both interpretations are mathematically equivalent.) Under this new light, $n$ becomes what Lindley (1957) calls "a measure of lack of conviction about the null hypothesis"

[2]Improper priors are extensions of the standard probability measures on the parameter space to infinite mass positive measures in order to reach more procedures and to close the inferential scope in several senses, see, e.g., Robert (2001).

[3]In terms of de Finetti's imaginary observations, the prior corresponds to the information brought by $n$ less *imaginary observations* than the real observations.

(p.189), a sentence that I re-interpret as the prior (under $H_1$) getting more and more diffuse as $n$ grows. However, I want to stress here that nowhere in Lindley's (1957) paper (nor in Jeffreys' (1939) book) is the difficulty with improper priors discussed.

In this (re-)interpretation of the Jeffreys–Lindley's paradox, I consider that the phenomenon exhibited therein is not paradoxical in the least: when the diffuseness of the (alternative) prior (i.e., under $H_1$) increases, the only relevant piece of information becomes that $\theta$ could be equal to $\theta_0$, to the extent that it overwhelms any evidence to the contrary contained in the data. For one thing, and as put by Lindley (1957), "the value $\theta_0$ is fundamentally different from any value of $\theta \neq \theta_0$, however near $\theta_0$ it might be" (p.189).[4] In addition, letting $n$ grow to infinity means that the mass of the prior distribution in any fixed neighbourhood of the null hypothesis and even in any set coherent with the data at hand vanishes to zero. There is therefore a deep coherence in the selection of the null hypothesis $H_0$ in this case: being completely indecisive about the alternative hypothesis means we could and should not chose this alternative. It is not possible to pick the alternative hypothesis of an undefined value of $\theta$ when opposed to the very special value $\theta_0$ if we want to be "completely non-informative" about $\theta$ under $H_1$. This analysis of the Jeffreys–Lindley's paradox is justifying (further) the prohibition of the use of improper priors for testing point null hypotheses and selecting embedded models (found for instance in DeGroot, 1982, Berger, 1985, and Robert, 2011).

Depending on one's perspective about the position of Bayesian statistics within statistical theories of inference, one might see this as a strength or as a weakness since Bayes factors and posterior probabilities do require a realistic model under the alternative when $p$-values and Bayesian predictives do not. One logical reason for this requirement is that Bayesian inference proceeds with the alternative model when the null is rejected.

## 3. DON'T BE AFRAID...

Under the rather provocative title *"Who should be afraid of the Jeffreys-Lindley paradox"*,[5] Spanos (2013) offers his frequentist reassessment of the paradox, arguing against both Bayesian and likelihood ratio approaches and in favour of the postdata severity evaluation he and Mayo have both been advocating since 2004. Analysing the criticisms contained in this paper of Spanos' towards a rebuttal of his theses is the primary reason for writing the current paper.

While I hope the reader has already gotten familiar with the contents of Spanos (2013), let me first recapitulate the main points made in this paper before embarking onto a more detailed analysis of those arguments. Spanos (2013) compare the frequentist (use of $p$-values), Bayesian (use of Bayes factors) and "likelihoodist" (use of likelihood ratios) approaches to statistical testing, with the conclusion that the latter two "give rise to highly fallacious results" (p.75), while the $p$-value can be processed (or saved) by the post-data severity analysis of Mayo and Spanos (2004), escaping the Jeffreys–Lindley's paradox paradox. The paper insists on the

[4] We will get return to this fundamental remark in the discussion of Spanos (2013) in the next section.

[5] Given the contents of the paper, the author presumably intends Bayesian statistics or Bayesians as the recipients of this question.

ability of this method to exhibit a certain degree $\gamma$ of discrepancy from the null hypothesis, while Bayesian and likelihoodist methods cannot and do not provide evidence for a particular alternative hypothesis (see, e.g., p.79). Spanos (2013) concludes that the paradox "has played an important role in undermining the credibility of frequentist inference" (p.91) as being "vulnerable to the fallacy of rejection" (p.91) but that the Bayes factor falls to the "fallacy of acceptance".[6]

### 3.1 A proper decisional framework for testing

The notion of *evidence* brought by the data in favour or against an hypothesis $H_0$ is never defined by Spanos (2013), even though it is repeatedly mentioned throughout the paper. More importantly, there is no argument made therein as to what the specific purpose of conducting a test (against, say, constructing a confidence interval) is. Spanos (2013) operates as though there were an obvious truth ($H_0$ or $H_1$) and as though one and only one statistical approach could reach it, despite the evidence to the contrary represented by the consistency property of all three approaches in Lindley's (1957) setting.[7]

Indeed, what differentiates statistical tests from other aspects of statistical inference is that (a) there is a precise question being asked about the statistical model under study and (b) the answer to this question will impact the subsequent actions of the individual who asked the question. Point (a) relates to Lindley's (1957) stress on the fact that $\theta_0$ is very special indeed and quite different from any neighbouring value: $\theta_0$ was select for a reason and with a motive, brought forward by a theoretical construct prior to observation rather than suggested from the data. From a Bayesian perspective, this specificity implies that prior information is available (to a certain degreee) as to why $\theta_0$ is a special value of the parameter $\theta$. Point (b) is about assessing the consequences of the answer to the questions, especially the wrong answer. Both from a frequentist and from a Bayesian perspective, this implies defining a loss or utility function that quantifies the impact of a wrong answer and eventually determines the boundary between acceptance and rejection.[8]

Spanos (2013) does not follow this decisional approach (which he considers as a Trojan horse for validating Bayesian inference, see Spanos, 2012). As shown by the remark "the problem does not lie with the $p$-value or the accept/reject rules as such, but with how such results are transformed into evidence for or against $H_0$ or a particular alternative" (p.76), the error statistical approach he advocates (as discussed in Section 3.3, it does not proceed from a decisional step, even when handling an accept/reject outcome), but instead requires the introduction of a secondary $p$-value bound, the *severity evaluation*, coupled with a parameter value (or deviation) that represents a significant distance from the null. *In fine*, this interpretation of testing relies on the use of an implicit loss function that sets

[6]The *fallacy of rejection* is "(mis)interpreting reject $H_0$ (evidence against $H_0$) as evidence for a particular $H_1$" (p.75), by which I understand for a specific value of the parameter under the alternative hypothesis, and the *fallacy of acceptance* is "(mis)interpreting accept $H_0$ (no evidence against $H_0$) as evidence for $H_0$" (p.75).

[7]Ironically, the numerical example used in the paper (borrowed from Stone, 1997, also father to the marginalisation paradoxes, see Dawid et al., 1973) is the very same as Bayes's billiard example (if with a larger value of $n$) and as Laplace's example on births (with a similar value of $n$).

[8]This is the simplest type of loss function: more advanced versions could include the case of a non-decision, calling for more observations, as in Berger (2003).

what is far and what is not. For instance, when Spanos (2013, p.75) states that "there is nothing fallacious or paradoxical about a small $p$-value or a rejection of the null, for a given significance level $\alpha$; when $n$ is large enough, since a highly sensitive test is likely to pick up on tiny (in a substantive sense) discrepancies from $H_0$", the "substantive sense" can only be gathered from a loss function. In connection with this notion of loss and of distance from the null hypothesis, Spanos' (2013) side remark that "what goes wrong is that the Bayesian factor and the likelihoodist procedures use Euclidean geometry to evaluate evidence for different hypotheses when in fact the statistical testing space is curved" (p.90) carries little weight. First, it is mathematically incorrect given that the Bayes factor is invariant under one-to-one reparameterisations of either the parameter or the sampling spaces, hence impervious to the curvatures of those spaces[9] and to the choice of a specific geometry. Second, the severity alternative put forward by Spanos in this paper rests upon the choice of a divergence measure $d(\mathbf{X})$ which is most often Euclidean, while the Bayesian and likelihood approaches rely on the likelihood function, which does not rely on the choice of a (Euclidean or not) distance.

### 3.2 The paradox as an anti-Bayesian argument

Spanos (2013) argues that the Jeffreys-Lindley's paradox is demonstrating against the Bayesian (and likelihood) resolutions of the problem by failing to account for the large sample size.[10] As detailed in Section 2.2, I do not disagree with this perspective to some extent: as discussed above, I consider that the most important lesson learned from Lindley (1957) is that improper priors require special caution when conducting point-null hypothesis testing. There is indeed little sense in arguing in favour of a procedure that would always conclude by picking the null, no matter what the value of the test statistics is. However, as pointed out in Section 2.1, considering a fixed (in the sample size $n$) value of the $t$ statistic has little meaning in an asymptotic referential, i.e. when $n$ increases to $\infty$. Either the $t$ statistic converges in distribution to the standard normal distribution under the null hypothesis $H_0$ or it diverges to infinity under the alternative $H_1$. This is the reason why both the Bayesian and the likelihood ratio approaches are consistent in this setting.[11]

In an encompassing perspective about hypothesis testing, I do argue that the Jeffreys-Lindley's paradox expresses foundational difficulties *for all* of the three methodological threads discussed in Spanos (2013): when following Fisher's approach, there is a theoretical and practical difficulty as to how one should decrease the acceptance bound $\alpha = \alpha(n)$ on the $p$-value when $n$ increases. This approach fails to provide a working and logical principle from which this bound (or sequence of bounds) $\alpha(n)$ should be chosen. For instance, the paper objects (p.78)

[9]Spanos (2013, p.90 and p.91) uses the term "statistical space" without a proper definition. It can be either the parameter or the sample space since there is no decision space in his axioms.

[10]His argument about the invariance of the Bayes factor to $n$ (p.84) is found missing as the Bayes factor does depend on $n$ as exhibited by $\mathfrak{B}(t_n)$ above.

[11]Somewhat in connection with this point, I fail to understand why a Bayes factor would "ignore the sampling distribution (...) by invoking the likelihood principle" (p.90): the Bayes factor incorporates the sampling distribution by integrating it out against the associated prior under the alternative hypothesis. There is no invoking involved and no likelihood principle at play in the construction of the marginal likelihood, but solely an application of the rule of probability calculus.

that because "of the large sample size, it is often judicious to choose a small type I error, say $\alpha = .003$" when this argument simply points at the arbitrariness of this numerical value. In the specific setting of this example of Spanos (2013), it could be worse, in that it could have been dictated by the data since the observed $p$-value takes the nearby .0027 value. In addition, I find the argument of consistence unconvincing in that case since both the Bayes factor and the likelihood ratio tests are then consistent testing procedures.

In the Neyman–Pearson referential, there is a fundamental difficulty in finding a proper balance (or imbalance) between Type I and Type II errors, since such balance is not provided by the theory, which settles for the sub-optimal selection of a *fixed* Type I error. In addition, the whole notion of *power*, while central to this referential, has arguable foundations in that this is a *function* that inevitably depends on the unknown parameter $\theta$. In particular, the power decreases to the Type I error at the boundary between the null and the alternative hypotheses in the parameter set. For instance, referring to Spanos' (2013) arguments, giving a meaning to the definition of severity (eqn. (25), p.87)

$$\mathbb{P}(\mathbf{x}; d(\mathbf{X}) < d(\mathbf{x}_0); \theta > \theta_1 \text{ is false}),$$

where $x_0$ is the observable and $\mathbf{x}$ should be $\mathbf{X}$, seems impossible. The third argument "$\theta > \theta_1$ is false" that conditions this probability statement makes no sense without a prior distribution on the parameter set.[12] Even the corrected version

$$\mathbb{P}_{\theta_1}(d(\mathbf{X}) < d(\mathbf{x}_0))$$

depends on the choice of the particular alternative $\theta_1$ or has to been seen as the power function which, like the risk function (see, e.g., Berger, 1985), prohibits most comparisons in a frequentist framework.

As discussed further in Section 2.2, apart from the genuine difficulty in setting a prior distribution over two distinct parameter spaces, following a standard Bayesian approach with a flat prior on the binomial probability inferred about in Spanos (2013) leads to a Bayes factor of 8.115 (p.80). Since this is neither a huge nor a tiny quantity *per se*, the very difficulty is in calibrating it, Jeffreys's (1939, Appendix) scale being highly formal.

### 3.3 Defending severity

Spanos (2013) rebounds on the failures (or fallacies?) of all three main approaches to address the difficulties with the Jeffreys–Lindley's paradox to advocate his own criterion the "postdata severity evaluation" introduced in an earlier paper with Deborah Mayo (Mayo and Spanos, 2004).[13] The notion of severe tests has been advocated by Mayo and Spanos (2004) over the past years, but it has

[12] An exchange with D. Mayo (2013, personal communication) led me to conclude that this probability is computed under the distribution of $\mathbf{X}$ associated with the parameter $\theta_1$, where $\theta_1$ is determined by the severity criterion, detailed in Section 3.3.

[13] Section 6 starts with the mathematically incorrect argument that, since we have observed $x_0$, in connection with the null hypothesis $H_0 :\ \theta = \theta_0$, the sign of $x_0 - \theta_0$ "indicates the relevant direction of departure from $H_0$". First, random variables may take values both sides of $\theta_0$ for most values of $\theta$. Second, the fact that one is testing $H_0$ against a two-sided or a one-sided alternative hypothesis pertains to the motivation of the test, not to the direction suggested by the data. The contentious modification of the testing setting *once* the data is observed is a major issue with Spanos' (2013) perspective, issue that we will discuss further below.

not yet made a dent on the theory or on the practice of statistical testing. As examplified by the paper (see, e.g., Table 1 on p.88 and the discussion surrounding it), this solution requires more (user-based) calibration than the regular $p$-value and it is thus bound to confuse practitioners. Indeed, the severity evaluation as explained[14] in Spanos (2013) implies defining for each departure from the null, rewritten as $\theta_1 = \theta_0 + \gamma$, the probability that a dataset associated with this parameter values "accords less with $\theta > \theta_1$ than $x_0$ does" (p.87). (Note that, as discussed in footnote 13, the two-sided alternative has been turned *postdatum* into a one-sided version. This is no more acceptable than stating that the date always supports more the value at the maximum likelihood then at the null.)

The notion of severity is therefore a mix of $p$-value and of Type II error that is supposed to "provide the 'magnitude' of the warranted discrepancy from the null" (p.88), i.e. to decide about how close (in distance) to the null we can get and still be able to discriminate the null from the alternative hypotheses "with very high probability" (p.86). The description found in Section 6 of Spanos (2013) implies a rejection of $H_0$ for the data at hand, based on the comparison of the $p$-value with an acceptance bound, as in Fisher's perspective, followed by an assessment of "the largest discrepancy $\gamma$ from $H_0$ warranted by data $x_0$", which derives a boundary parameter value $\theta_1$ from a severity level, .9 say. This amounts to selecting a minimal power or maximal Type II error and to check for the corresponding discrepancy to be "substantively significant" (p.88), an assessment provided by the user and thus defeating the original purpose of the approach. The error statistical approach is therefore ineffective as an operational tool, that is, in practice, because of this double calibration that is required from the user. Once more, as detailed in Spanos (2013), the value of this closest discrepancy $\gamma$—which is thus a bound on where we can discriminate between $H_0$ and $H_1$ for a given sample size $n$—does depend on another arbitrary tail probability, the "severity threshold",

$$\mathbb{P}_{\theta_1}\{d(\mathbf{X}) \leq d(x_0)\}$$

introduced in eqn. (25). This tail probability has to be chosen by the user without being more intuitive or less subjective than the initial acceptance bound on the $p$-value.[15] Furthermore, once the resulting discrepancy $\gamma$ is found, whether it is far enough from the null is a matter of informed opinion since, as duly noted by Spanos (2013), whether it is "substantially significant (...) pertains to the substantive subject matter" (p.88), implying once again some sort of loss function or of prior information that the paper fails to acknowledge.[16]

[14]Typos in both the last line in p.87, which is mixing the standardised and the non-standardised versions of the test statistic, and Table 1, which introduces a superfluous minus sign, do not help in clarifying the issue.

[15]When considering the severity as a function of $\theta_1$, complement to a probability cdf in $\theta_1$, its most natural interpretation would be of a Bayesian nature, the bound being then a prior quantile. However, this solution is quite improbable to meet with the authors' approval.

[16]While this is very much unlikely to be advocated either by the author or by Bayesian statisticians, we note that, as a statistics, i.e. a transform of the data, both the Bayes factor and the likelihood ratio could be processed in exactly the same way to produce severity thresholds of their own.

### 3.4 Falling afoul of the fallacy of rejection

In connection with the special meaning of the value $\theta_0$ and with the argument of the fallacy of rejection, mentioned by Spanos (2013) as associated with the $p$-value, several parts of his discussion of the Bayesian approach argue (see, e.g., p.81) that other values of $\theta$ are supported and even *better* supported by the data than the null value $\theta_0$. This is a surprising argument as (a) it pertains to the construction of Bayesian credible intervals but not to testing and (b) it is a direct illustration of the "fallacy of rejection" in that rejecting (or not) $H_0$ does not bring evidence in favour of a particular value of $\theta$. While it is correct that the observed data $\mathbf{x}_0$ does "favor certain values [of the parameter] more strongly" (p.81) than $\theta_0$, those values are (a) driven by the data, i.e. will vary from one repetition of the statistical experiment to the next, and (b) of no particular relevance for conducting a test, meaning that the experimenter or the scientist behind the experiment had not expressed a particular interest in those values before they were exposed by the data. The tested value, $\theta_0 = 0.2$ say, is chosen prior to the experiment because it has some special meaning for the problem or the theory at hand. The fact that the likelihood and/or the posterior are/is larger in other values of $\theta$ does not constitute "conflicting evidence" (p.82) against the fact that the null hypothesis holds. It simply reflects on the property that the likelihood function is a random function of the parameter $\theta$, whose mode also varies with the data and is almost surely not located at the true value of the parameter. Since this is a mathematical evidence, I find astounding that it can be used as a logical argument against some statistical approaches to testing.

## 4. TOWARD A RESOLUTION OF THE BAYESIAN VERSION OF THE PARADOX

While the divergence between the frequentist and Bayesian answers is reflecting upon the difference between the paradigms in terms of purpose and evaluation, rather than condemning one of the approaches as implied by Spanos (2013), the (Bayesian) debate about constructing limiting Bayes factors or posterior probabilities that include improper prior modelling stands both open and relevant. DeGroot's (1982) warning that "diffuse prior distributions (...) must be used with care" has now been impressed upon generations of students and it is indeed a fair warning. There remains nonetheless a crucial need to produce assessments of null hypotheses from a Bayesian perspective and under limited prior information, once again without any incentive whatsoever to mimic, reproduce or even come close to frequentist solutions like $p$-values.

In Robert (1993), I suggested selecting the prior weights of the two hypotheses, $(\varrho_0, 1-\varrho_0)$, in order to compensate for the increased mass produced by the alternative hypothesis prior.[17] While the solution therein produced numerical results that brought a proximity with the $p$-value, its construction is flawed from a measure-theoretic point of view since the determination of the weights involves the value of the prior density $\pi_1$ at the point-null value $\theta_0$,

$$\varrho_0 = (1-\varrho_0)\pi_1(\theta_0)\,,$$

[17]The compensation cannot be probabilistic in that the overall mass of an improper prior remains improper for any weighting scheme.

a difficulty also shared by the (related) Savage–Dickey paradox (Robert and Marin, 2009).[18] I nonetheless remain of the opinion that the degree of freedom represented by the prior weight $\varrho_0$ in the Bayesian formalism should not be neglected to overcome the difficulty in using improper priors.[19]

A further step worth mentioning is Berger et al.'s (1998) partial validation of the use of *identical* improper priors on the nuisance parameters, a notion already entertained by Jeffreys (1939, see the discussion in Robert et al., 2009, Section 6.3). While arguing about the case of the"same" constant in both models as validating picking the "same" improper prior for both models has neither mathematical nor statistical validation, relying on the same prior quite handily eliminates the major thorn in the side of Bayesian testing of hypotheses. As demonstrated in Marin and Robert (2007) and Celeux et al. (2012), it allows in particular for the use of a partly improper $g$-prior in linear and generalised linear models (Zellner, 1986).[20]

A last step towards the incorporation of improper priors within the Bayesian testing paraphenalia is the recent investigation of the use of score functions $S(x, m)$ that extend the standard log score function associated with the Bayes factor:

$$\log B_{12}(x) = \log m_1(x) - \log m_2(x) = S_0(x, m_1) - S_0(x, m_2) \,,$$

where $m_i$ is the prior predictive associated with model $\mathfrak{M}_i$. Indeed, there exists a whole family of proper scoring rules that are independent from the normalising constant of the prior predictive (Parry et al., 2012) and can thus be used on improper priors as well. For instance, Hyvärinen's (2005) score is one of these scores. While the scores are delicate to calibrate, i.e. the magnitude of $S(x, m_1) - S(x, m_2)$ is not absolute, they provide a consistent method for selecting models (Berger, 1985) and avoid the delicate issue of selecting priors that differ for model selection and for regular inference (conditional on the model). This is why Sprenger (2013) advises replacing the Bayes factor with a logarithmic score, rewritten as

$$\mathbb{E}^\pi \left[ \mathbb{E}_\theta \{ \log f(X|\theta)/f(X|\theta_0) \} | x \right] \,,$$

and compared with an acceptance bound. The Kullback–Leibler divergence used in this score is utterly natural in terms of evaluating the impact of replacing one distribution with the other. And, as stressed by Sprenger (2013), it does not "involve commitment to the truth or likelihood of $H_0$". The use of this score however requires the choice of an acceptance bound, which calibration is not provided by the theory.

[18] A solution to the measure-theoretic difficulty is to impose a version of $\pi_1$ that is continuous at $\theta_0$ so that $\pi_1(\theta_0)$ is uniquely defined. It however equates the values of two density functions under two orthogonal measures.

[19] Some will object at this choice on Bayesian grounds as it implies that the prior does depend on the sample size $n$.

[20] Once again, choosing $g = n$ should attract criticism from some Bayesian corners for being dependent on the sample size, even though it boils down to picking an imaginary sample (Smith and Spiegelhalter, 1982) size of 1. See Liang et al. (2008) for an alternative approach setting an hyperprior on $g$.

## 5. REFLECTIONS

The appeal of great paradoxes[21] is to address foundational issues in a field, either to reinforce the arguments in favour of a given theory or, on the opposite, to cast serious doubts on its validity. The fact that the Jeffreys–Lindley's paradox is still discussed in papers (as exemplified by Spanos, 2013 and Sprenger, 2013 in the current journal) and blogs, by statisticians and non-statisticians alike, is a testimony to its impact on the debate about the deepest foundations of statistical testing. The irrevocable opposition between frequentist and Bayesian approaches to testing, but also the persistent impact of the prior modelling in this case, are fundamental questions that have not yet met with definitive answers. And they presumably never will for, as aptly put by Lad (2003), "the weight of Lindley's paradoxical result (...) burdens proponents of the Bayesian practice". However, this is a burden with highly positive features in that it paradoxically drives the field to higher grounds.[22]

## REFERENCES

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. Springer-Verlag, New York.

Berger, J. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, **18** 1–32.

Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection* (P. Lahiri, ed.), vol. 38 of *Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, Beachwood Ohio, 135–207.

Berger, J., Pericchi, L. and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya A*, **60** 307–321.

Berger, J. and Sellke, T. (1987). Testing a point-null hypothesis: the irreconcilability of significance levels and evidence (with discussion). *J. American Statist. Assoc.*, **82** 112–122.

Camus, A. (1942). *Le Mythe de Sisyphe*. Gallimard, Paris.

Celeux, G., Anbari, M. E., Marin, J. and Robert, C. (2012). Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, **7** 477–502.

Dawid, A., Stone, N. and Zidek, J. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Royal Statist. Society Series B*, **35** 189–233.

DeGroot, M. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *J. American Statist. Assoc.*, **68** 966–969.

DeGroot, M. (1982). Discussion of Shafer's 'Lindley's paradox'. *J. American Statist. Assoc.*, **378** 337–339.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis* (revised edition). Chapman & Hall/CRC, New York.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, **6** 695–709. (electronic).

Jeffreys, H. (1939). *Theory of Probability*. 1st ed. The Clarendon Press, Oxford.

Kass, R. and Wasserman, L. (1996). Formal rules of selecting prior distributions: a review and annotated bibliography. *J. American Statist. Assoc.*, **91** 343–1370.

Lad, F. (2003). Appendix: the Jeffreys–Lindley paradox and its relevance to statistical testing. Tech. rep., Conference on Science and Democracy, Palazzo Serra di Cassano, Napoli.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation* (revised edition). Springer-Verlag, New York.

Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2008). Mixtures of $g$ priors for Bayesian variable selection. *J. American Statist. Assoc.*, **103** 410–423.

[21] I use this term despite my reluctance to call such phenomena "paradoxes", since they correspond to neither logical impossibilities nor to mathematical mistakes, but rather to contradictions in reasoning or, in the current case, to attempts to bring two different paradigms together.

[22] To conclude with a literary quote, "il faut imaginer Sisyphe heureux" (Camus, 1942).

Lindley, D. (1957). A statistical paradox. *Biometrika*, **44** 187–192.

Lindley, D. (1991). Discussion of the paper by Aitkin. *J. Royal Statist. Society Series B*, **53** 130–131.

Marin, J. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York.

Mayo, D. and Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, **71** 1007–1725.

Parry, M., Dawid, A. and Lauritzen, S. (2012). Proper local scoring rules. *Ann. Statist.* 561–592.

Robert, C. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica*, **3** 601–608.

Robert, C. (2001). *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York.

Robert, C., Chopin, N. and Rousseau, J. (2009). Theory of Probability revisited (with discussion). *Statist. Science*, **24(2)** 141–172 and 191–194.

Robert, C. and Marin, J.-M. (2009). On resolving the Savage–Dickey paradox. *Elect. J. Statistics*.

Shafer, G. (1982). On Lindley's paradox (with discussion). *J. American Statist. Assoc.*, **378** 325–351.

Smith, A. and Spiegelhalter, D. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Royal Statist. Society Series B*, **44** 377–387.

Spanos, A. (2012). Why the decision theoretic perspective misrepresents frequentist inference: 'Nuts and bolts' vs. learning from data. Technical report, arxiv:1211.0638.

Spanos, A. (2013). Who should be afraid of the Jeffreys–Lindley paradox? *Philosophy of Science*, **80** 73–93.

Spreger, J. (2013). Testing a precise null hypothesis: The case of Lindley's paradox *Philosophy of Science*, to appear.

Stone, M. (1997). Discussion of Aitkin (1997). *Statistics and Computing*, **7** 263–264.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with $g$-prior distribution regression using Bayesian variable selection. In *Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti*. North-Holland / Elsevier, 233–243.