Testing a Precise Null Hypothesis: The Case of Lindley's Paradox

Jan Sprenger*

July 13, 2012

Abstract

The interpretation of tests of a point null hypothesis against an unspecified alternative is a classical and yet unresolved issue in statistical methodology. This paper approaches the problem from the perspective of Lindley's Paradox: the divergence of Bayesian and frequentist inference in hypothesis tests with large sample size. I contend that the standard approaches in both frameworks fail to resolve the paradox. As an alternative, I suggest the Bayesian Reference Criterion: (i) it targets the predictive performance of the null hypothesis in future experiments; (ii) it provides a proper decision-theoretic model for testing a point null hypothesis and (iii) it convincingly accounts for Lindley's Paradox.

*Contact information: Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

1 Introduction. Lindley's Paradox.

Lindley's Paradox is one of the most salient cases where subjective Bayesian and frequentist inference fall apart. The paradox emerges in statistical tests of point null hypotheses with high sample sizes.

Instead of starting with a theoretical definition of the paradox, we give an example with real data (Jahn, Dunne and Nelson 1987). The case at hand involved the test of a subject's claim to possess extrasensory capacities (ESP) that would enable him to affect a series of 0-1 outcomes generated by a randomly operating machine ($\theta_0 = 0.5$). The subject claimed that these capacities would make the sample mean differ significantly from 0.5.

The sequence of zeros and ones, $X_1, \ldots, X_N$, was described by a Binomial model $B(\theta, N)$. The null hypothesis asserted that the results were generated by a machine operating with a chance of $H_0 : \theta = \theta_0 = 1/2$, whereas the alternative was the unspecified hypothesis $H_1 : \theta \neq \theta_0$. The experimenters decided to observe a very long series of zeros and ones, which would give us enough evidence as to judge whether or not the null was compatible with the data.

Jahn, Dunne and Nelson (1987) report that in 104.490.000 trials, 52.263.471 ones and 52.226.529 zeros were observed. A classical, Fisherian frequentist would now calculate the z-statistic which is

$$z(x) := \sqrt{\frac{N}{\theta_0(1-\theta_0)}} \left(\frac{1}{N} \sum_{i=1}^{N} x_i - \theta_0 \right) \approx 3.61 \tag{1}$$

and reject the null hypothesis on the grounds of the very low *p-value* it induces:

$$p := P_{H_0}(|z(X)| \geq |z(x)|) \ll 0.01 \tag{2}$$

Thus, the data would be interpreted as strong evidence for extrasensory capacities. Compare this now to the result of a Bayesian analysis. Jefferys (1990) assigns a conventional positive probability $P(H_0) = \varepsilon > 0$ to the null hypothesis and calculates the *Bayes factor* in favor of the null (the ratio of prior and posterior odds):

$$B_{01}(x) := \frac{P(H_0|x)}{P(H_1|x)} \cdot \frac{P(H_1)}{P(H_0)} \approx 19$$

Hence, the data strongly favor the null over the alternative and do *not* provide evidence for the presence of ESP.

The divergence between Bayesians and frequentists can be generalized. Arguably, what is most distinctive about the above example is the large sample size. Now assume that we are comparing observation sets of different sample size N, all of which attain, in frequentist terms, the same p-value, e.g., the highly significant value of 0.01. This means that the standardized sample mean

$z(x) = \sqrt{N}(\bar{x} - \theta_0)/\sigma$ takes the same value for all observation sets, regardless of the actual sample size. However, in that case, the Bayesian evaluation of the data will become ever more inclined to the null hypothesis with increasing N. Thus, a result that speaks highly significantly against the null from a frequentist point of view can strongly support it from a Bayesian perspective. This problem has, since the seminal paper of Lindley (1957), been known as *Lindley's Paradox*.

Due to its prominence and its simplicity, Lindley's Paradox is a suitable test case for comparing various philosophies of statistical inference, and for reconsidering the goals and methods of testing a precise null hypothesis. In this paper, I ask the following questions: First, which statistical analysis of the ESP example is correct? Second, which implications has Lindley's Paradox for standard procedures of Bayesian and frequentist inference? Third, is there a full decision-theoretic framework in which point null hypothesis tests can be conducted without adopting a fully subjectivist perspective? I will argue that both the standard Bayesian and the standard frequentist way to conceive of Lindley's Paradox are unsatisfactory, and that alternatives have to be explored. In particular, I believe that José Bernardo's Bayesian Reference Criterion holds considerable promise as a replication-oriented decision model that fits our intuitions about Lindley's Paradox.

2 Testing a precise null: frequentist vs. Bayesian accounts

Lindley's Paradox deals with tests of a precise null hypothesis $H_0 : \theta = \theta_0$ against an unspecified alternative $H_1 : \theta \neq \theta_0$ for large sample sizes. But why are we actually testing a precise null hypothesis if we know in advance that this hypothesis is, in practice, never *exactly* true? (For instance, in tests for the efficacy of a medical drug, it can safely be assumed that even the most unassuming placebo will have some minimal effect, positive or negative.)

The answer is that precise null hypotheses give us a useful idealization of reality for the purpose at hand. This is also rooted in Popperian philosophy of science: "only a highly testable or improbable theory is worth testing and is actually (and not only potentially) satisfactory if it withstands severe tests" (Popper 1963, 219–220). Accepting such a theory is not understood as endorsing the theory's truth, but as choosing it as a guide for future predictions and theoretical developments.

Frequentists have taken the baton from Popper and explicated the idea of severe testing by means of statistical hypothesis tests. Their mathematical rationale is that if the discrepancy between data and null hypothesis is large enough, we can infer the presence of a significant effect and reject the null hypothesis. For measuring the discrepancy in the data $x := (x_1, \ldots, x_N)$ with respect to postulated mean value θ_0 of a Normal model, one canonically uses the standardized statistic

$$z(x) := \frac{\sqrt{N}}{\sigma}\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0\right)$$

that we have already encountered above. Higher values of z denote a higher divergence from the null, and vice versa. Since the distribution of z usually varies with the sample size, some kind of standardization is required. Many practitioners use the *p-value* or *significance level*, that is, the "tail area" of the null hypothesis under the observed data, namely $p := P_{H_0}(|z(X)| \geq |z(x)|)$.

On that reading, a low p-value indicates evidence against the null: the chance that z would take a value at least as high as $z(x)$ is very small, if the null were indeed true. Conventionally, one says that $p < 0.05$ means significant evidence against the null, $p < 0.01$ very significant evidence, or in other words, the null hypothesis is rejected at the 0.05 level, etc. R.A. Fisher has interpreted p-values as "a measure of the rational grounds for the *disbelief* [in the null hypothesis] it augments" (Fisher 1956, 43).

Subjective Bayesians choose a completely different approach to hypothesis testing. For them, scientific inference obeys the rules of probabilistic calculus. Probabilities represent honest, subjective degrees of belief, which are updated by means of Bayesian Conditionalization. A Bayesian inference about a null hypothesis is based on the posterior probability $P(H_0|E)$, the synthesis of data E and prior $P(H_0)$.

It is here that Bayesians and significance testers clash with each other. If the p-value is supposed to indicate to what extent the null is still tenable, we get a direct conflict with Bayesian reasoning. The analyses of Berger and Delampady (1987) and Berger and Sellke (1987) show that p-values tend to grossly overstate evidence against the null, to the extent that the posterior probability of the null – and even the *minimum* of $P(H_0|x)$ under a large class of priors – is typically much higher than the observed p-value. In other words, even a Bayesian analysis that is maximally biased against the null is still less biased than a p-value analysis. This has led Bayesian statisticians to conclude that "almost anything will give a better indication of the evidence provided by the data against H_0" (Berger and Delampady 1987, 330). These findings are confirmed by methodologists in the sciences who have repeatedly complained about the illogic of p-values (and significance testing) and their inability to answer the questions that really matter for science (Cohen 1994; Royall 1997; Goodman 1999).

Lindley's Paradox augments this divergence of a Bayesian and a frequentist analysis. In a Normal model, if $P(H_0) > 0$ and $N \to \infty$, then the posterior probability of the null $P(H_0|x)$ converges to 1 for almost any prior distribution over H_1. More precisely:

> **Lindley's Paradox:** Take a Normal model $N(\theta, \sigma^2)$ with known variance σ^2, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, assume $P(H_0) > 0$ and any regular proper prior distribution on $\{\theta \neq \theta_0\}$. Then, for any testing level $\alpha \in [0, 1]$, we can find a sample size $N(\alpha)$ and independent, identically distributed data $x = (x_1, \ldots, x_N)$ such that
>
> 1. The sample mean $\bar{x}$ is significantly different from θ_0 at level α;
> 2. $P(H_0|x)$, that is, the posterior probability that $\theta = \theta_0$, is at least as big as $1 - \alpha$. Lindley (cf. 1957, 187)

One might conjecture that this Bayesian-frequentist divergence stems from the unrealistic assumption that $P(H_0) > 0$. But actually, the findings are confirmed if we switch to an analysis in terms of Bayes factors, the Bayesian's standard measure of evidence. The evidence x provides for H_0 vis-à-vis H_1 is written as B_{01} and defined as the ratio of prior and posterior odds:

$$B_{01}(x) := \frac{P(H_0|x)}{P(H_1|x)} \cdot \frac{P(H_1)}{P(H_0)} = \frac{P(x|H_0)}{P(x|H_1)}, \tag{3}$$

which can alternatively be interpreted as an averaged likelihood ratio of H_0 vs. H_1. Now, if the prior over H_1, that is, the relative weight of alternatives to the null, follows a $N(\theta_0, \tilde{\sigma}^2)$-distribution, then the Bayes factor in favor of the null can be computed as

$$B_{01}(x) \quad = \quad \sqrt{1 + \frac{N\tilde{\sigma}^2}{\sigma^2}}\, e^{\frac{-Nz(x)^2}{2N+2\sigma^2/\tilde{\sigma}^2}}, \tag{4}$$

which converges, for increasing N, to infinity as the second factor is bounded (Bernardo 1999, 102). This demonstrates that the precise value of $P(H_0)$ is immaterial for the outcome of the subjective Bayesian analysis.

This result remarkably diverges from the frequentist finding of significant evidence against the null. What has happened? If the p-value, and consequently the value of $z(X) = c$, remain constant for increasing N, we can make use of the Central Limit Theorem: $z(X)$ converges, for all underlying distributions with bounded second moments, in distribution against $N(0, 1)$. Thus, as $N \to \infty$, we obtain that $c\sigma \approx \sqrt{N}(\bar{X} - \theta_0)$, and $\bar{X} \to \theta_0$. In other words, the sample mean gets ever closer to θ_0, favoring the null over the alternatives. For the deviance between the variance-corrected sample mean z and H_0 will be relatively small compared to the deviance between z and all those hypotheses in H_1 that are "out there", in sharp contrast to a frequentist tester who will observe significant evidence against H_0.

In other words: as soon as we take our priors over H_1 seriously, as an expression of our uncertainty about which alternatives to H_0 are more likely than others, we will, in the long run, end up with results favoring θ_0 over an unspecified alternative. Bayesians read this as the fatal blow for frequentist inference since an ever smaller deviance of the sample mean $\bar{x}$ from the parameter value θ_0 will suffice for a highly significant result. Obviously, this makes no scientific sense. Small, uncontrollable biases will be present in any record of data, and frequentist hypothesis tests are unable to distinguish between *statistical significance* ($p < 0.05$) and *scientific significance* (a real effect is present). A Bayesian analysis, on the other hand, accounts for this insight: as $\bar{X} \rightarrow \theta_0$, an ever greater chunk of the alternative H_1 will diverge from $\bar{X}$, favoring the null hypothesis.

Still, the subjective Bayesian stance on hypothesis tests leaves us with an uneasy feeling. Assigning a strictly positive degree of belief $P(H_0) > 0$ to the point null hypothesis $\theta = \theta_0$ is a misleading and inaccurate representation of our subjective uncertainty. In terms of degrees of belief, θ_0 is not that different from any value $\theta_0 \pm \varepsilon$ in its neighborhood. Standardly, we would assign a continuous prior over the real line, and there is no reason why a set of measure zero, namely $\{\theta = \theta_0\}$, should have a strictly positive probability. But if we set $P(H_0) = 0$, then for most priors (e.g., an improper uniform prior) the posterior probability distribution will not peak at the null value, but somewhere else. Thus, the apparently innocuous assumption $P(H_0) > 0$ has a marked impact on the result of the Bayesian analysis.

A natural reply to this objection contends that H_0 is actually an idealization of the hypothesis $|\theta - \theta_0| < \epsilon$, for some small ϵ, rather than a precise point null hypothesis $\theta = \theta_0$. Then, it would make sense to use strictly positive priors. Indeed, it has been shown that point null hypothesis tests in terms of Bayes factors approximate a test of whether a small interval around the null contains the true parameter value (Theorem 1 in Berger and Delampady 1987). Seen that way, it *does* make sense to assign a strictly positive prior to H_0.

Unfortunately, this won't help us in the situation of Lindley's Paradox: when $N \rightarrow \infty$, the convergence results break down, and testing a point null is no more analogous to testing whether a narrow interval contains θ. In the asymptotic limit, the Bayesian cannot justify the strictly positive probability of H_0 as an approximation to testing the hypothesis that the parameter value is close to θ_0 – which is the hypothesis of real scientific interest. Setting $P(H_0) > 0$ may be regarded as a useful convention, but this move neglects that a hypothesis test in science asks, in the first place, if H_0 is a reasonable simplification of a more general model, and not if we assign a high degree of belief to this precise value of θ.

This fact may be the real challenge posed by Lindley's Paradox. In the debate with frequentists, the Bayesian likes to appeal to "foundations", but working with strictly positive probabilities of the null hypothesis is hard to justify from a foundational perspective, and also from the perspective of scientific practice.

The bottom line of all is that the subjective Bayesian analysis fails to explain why hypothesis tests have such an appeal to scientific practitioners, and even to those that are statistically sophisticated. Similarly, the Bayesian has a hard time to explain why informative and precise, but improbable hypotheses should sometimes be preferred over more general alternatives. How can the subjectivist model that we are less interested in the *truth* of H_0 than in its *usefulness*?

3 The BRC approach to hypothesis testing

This section presents a proposal for a fully Bayesian decision model for hypothesis testing that survives the criticisms raised against the subjectivist approach and gives a satisfactory treatment of Lindley's Paradox. The main idea is to decouple the idea of testing a precise null hypothesis H_0 from the truth of this hypothesis. Instead, we view the statistical test as making a decision on whether or not we should treat the null hypothesis $H_0 : \theta = \theta_0$ as a proxy for the more general model $H_1 : \theta \neq \theta_0$. In other words, we test whether the null is compatible with the data using a specific utility structure, going back to the roots of Bayesianism in decision theory.

Thus, we have to extend Bayesian belief revision to Bayesian decision models and add a proper utility dimension. This allows for much more flexible treatments than the traditional zero-one loss model that is implicitly presupposed in inference to the most probable hypothesis. In the remainder, I sketch a simplified version of Bernardo's Reference Bayesian Criterion (1999, section 2-3) in order to elaborate the main ideas of philosophical interest.

In science, we generally prefer hypotheses on whose predictions we may rely. Therefore, a central component of the envisioned decision model depends on the expected predictive accuracy of the null. Hence, we need a function that evaluates the predictive score of a hypothesis, given some data y. The canonical approach consists in the logarithmic score $\log P(y|\theta)$ (Good 1952): if an event considered to be likely occurs, then the score is high; if an unlikely event occurs, the score is low. This is a natural way of rewarding good and punishing bad predictions.

A generalization of this utility function describes the score of data y under parameter value θ as $q(\theta, y) = \alpha \log P(y|\theta) + \beta(y)$, where α is a scaling term, and

$\beta(y)$ is a function that depends on the data only. Informally speaking, $q(\cdot,\cdot)$ is decomposed into a prediction-term and a term that depends on the desirability of an outcome, where the latter will eventually turn out to be irrelevant. This is a useful generalization of the logarithmic score. Consequently, if θ is the true parameter value, the utility of taking H_0 as a proxy for the more general model H_1 is

$$\int q(\theta_0, Y)\, dP_{Y|\theta} \quad = \quad \alpha \int \log P(y|\theta_0)\, P(y|\theta)\, dy + \int \beta(y)\, P(y|\theta)\, dy.$$

The overall utility U of a decision, however, should not only depend on the predictive score, as captured in q, but also on the cost c_j of selecting a specific hypothesis H_j. Ceteris paribus, H_0 should be preferred to H_1 because it is more informative, simpler, and less prone to the risk of overfitting (in case there are nuisance parameters). Therefore it is fair to set $c_1 > c_0$. Writing $U(\cdot,\theta) = \int q(\cdot, Y)\, dP_{Y|\theta} - c_j$, we then obtain

$$U(H_0,\theta) = \alpha \int \log P(y|\theta_0)\, P(y|\theta)\, dy + \int \beta(y)\, P(y|\theta) dy - c_0$$

$$U(H_1,\theta) = \alpha \int \log P(y|\theta)\, P(y|\theta)\, dy + \int \beta(y)\, P(y|\theta) dy - c_1.$$

Note that the utility of accepting H_0 is evaluated against the true parameter value θ, and that the alternative is not represented by a probabilistic average (e.g., the posterior mean), but by its best element, namely θ. This is arguably more faithful than subjective Bayesianism to the essential asymmetry in testing a point null hypothesis. Consequently, the difference in *expected utility*, conditional on the posterior density of θ, can be written as

$$\begin{aligned}
&\int_{\theta\in\Theta} \left(U(H_1,\theta) - U(H_0,\theta)\right)\, P(\theta|x)\, d\theta \\
= \quad & \alpha \int_{\theta\in\Theta} \left(\int \log \frac{P(y|\theta)}{P(y|\theta_0)}\, P(y|\theta) \right) P(\theta|x)\, dy\, d\theta + \int \beta(y)\, P(y|\theta)\, dy \\
& - \int \beta(y)\, P(y|\theta)\, dy + c_0 - c_1 \\
= \quad & \alpha \int_{\theta\in\Theta} \left(\int \log \frac{P(y|\theta)}{P(y|\theta_0)}\, P(y|\theta)\, dy \right) P(\theta|x)\, d\theta + c_0 - c_1.
\end{aligned}$$

This means that the expected utility difference between inferring to the null hypothesis and keeping the general model is essentially a function of the expected log-likelihood ratio between the null hypothesis and the true model, calibrated against a "utility constant" $d^*(c_0 - c_1)$. For the latter, Bernardo suggests a conventional choice that recovers the well-probed scientific practice of regarding three standard deviations as strong evidence against the null. The exact value of d^* depends, of course, on the context: on how much divergence is required to balance the advantages of working with a simpler, more informative, and more accessible model (Bernardo 1999, 108).

Wrapping up all this, we will reject the null if and only if $\mathbb{E}_\theta[U(H_1, \theta)] > \mathbb{E}_\theta[U(H_0, \theta)]$ which amounts to the

> **Bayesian Reference Criterion (BRC):** Data x are incompatible with the null hypothesis $H_0 : \theta = \theta_0$, assuming that they have been generated from the probability model $(P(\cdot|\theta), \theta \in \Theta)$, if and only if
>
> $$\int_{\theta \in \Theta} P(\theta|x) \left(\int \log \frac{P(y|\theta)}{P(y|\theta_0)} P(y|\theta)\, dy \right) d\theta > d^*(c_0 - c_1). \qquad (5)$$

This approach has a variety of remarkable features. First, it puts hypothesis testing on firm decision-theoretic grounds. Second, accepting the null, that is, using θ_0 as a proxy for θ, amounts to claiming that the difference in expected predictive success of θ_0 and the true parameter value θ will be offset by the fact that H_0 is more elegant, more informative and easier to test. Hence, BRC does not only establish a tradeoff between different epistemic virtues: it is also in significant agreement with Popper's view that "science does not aim, primarily, at high probabilities. It aims at high informative content, well backed by experience." (Popper 1934/59, 399). Third, the approach is better equipped than subjective Bayesianism to account for frequentist intuitions, since under some conditions, e.g., in Lindley's Paradox, the results of a reference Bayesian analysis agree with the results of a frequentist analysis, as we shall see below. Fourth, it is invariant of the particular parametrization, that is, the final inference does not depend on whether we work with θ or a 1:1-transformation $g(\theta)$. Fifth, it is neutral with respect to the kind of prior probabilities that are fed into the analysis.[1]

4 Revisiting Lindley's Paradox

We now investigate how Bernardo's approach deals with Lindley's Paradox and return to the ESP example from the introduction. It turns out that the BRC quantifies the expected loss from using θ_0 as a proxy for the true value θ as substantial. Using a $\beta(1/2, 1/2)$ reference prior for θ (Bernardo 1979), the expected loss under the null hypothesis is calculated as $d(\theta = 1/2) \approx \log 1400 \approx 7.24$. This establishes that "under the accepted conditions, the precise value $\theta_0 = 1/2$ is rather incompatible with the data" (Bernardo 2012, 18).

[1] BRC implies that some parameters, such as d^*, which have to be chosen conventionally or context-dependent. Hence, a charge of "arbitrariness" could be made. However, this flexibility is, in my opinion, an asset of a general decision-theoretic model, not a drawback, as a comparison with Expected Utility Theory makes clear.

We observe that the results of a reference analysis according to BRC agree with the results of the frequentist analysis, but contradict the subjective Bayesian results. One might thus object that Bernardo's Bayesianism is purely *instrumental*: that is, it makes use of Bayesian notation and assigns a "probability" over θ, but it ends up with conventional, automated inference procedures that recover frequentist results.

Let us get back to the experiment. Of course, the rejection of the null hypothesis does not prove the extrasensory capacities of our subject; a much more plausible explanation is a small bias in the random generator. This is actually substantiated by looking at the posterior distribution of θ: due to the huge sample size, we find that for any non-extreme prior probability function, we obtain the posterior $\theta \sim N(0.50018, 0.000049)$, which shows that most of the posterior mass is concentrated in a narrow interval that does *not* contain the null. These findings agree with a likelihood ratio analysis: if we compute the log-likelihood ratio $L_{\hat{\theta},\theta_0}$ of the maximum likelihood estimate $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$ versus the null, we obtain (using the Normal approximation)

$$\begin{aligned}
\log L_{\hat{\theta},\theta_0}(x_1, ..., x_N) &= \log \frac{P(\bar{x}|\hat{\theta})}{P(\bar{x}|\theta_0)} = \log \frac{P(x_1 = \hat{\theta}|\hat{\theta})^N}{P(x_1 = \hat{\theta}|\theta_0)^N} \\
&= \log \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N - \log \left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{N}{2\sigma^2}(\hat{\theta}-\theta_0)^2}\right) \\
&= \frac{N}{2\sigma^2}(\hat{\theta} - \theta_0)^2 \overset{N\to\infty}{\longrightarrow} \infty. \qquad (6)
\end{aligned}$$

This analysis clearly shows that the likelihood ratio with respect to the maximum likelihood estimate speaks, for large N, increasingly *against* the null (in our case: $\log L_{\hat{\theta},\theta}(x_1, ..., x_N) \approx 6.53$), in striking disagreement with the Bayes factor analysis.

If we revisit Jeffery's analysis in the light of these observations, we note two contentious features, already touched upon previously. The first concerns the utility structure that is imposed by basing inference exclusively on the posterior distribution. We have seen in the previous sections that such a zero-one loss function, and a positive prior probability $P(H_0)$ may not be adequate assumptions for deciding whether a hypothesis should be judged as compatible with the data; therefore we should also be wary of judgments based on such assumptions. Second, a Bayes factor comparison effectively compares the likelihood of the data under H_0 to the *averaged likelihood* of the data under H_1. However, this quantity is strongly influenced by whether there are some extreme hypotheses in H_1 that fit the data poorly. Compared to the huge amount of data that we have just collected, the impact of these hypotheses (mediated via the conventional uniform prior) should be minute. These arguments explain why most people would tend to judge the data as incompatible with the *precise* null, but fail to see a scientifically interesting effect.

From the vantage point of whether the experimental effect is likely to be *replicated* – and this is a question scientists are definitely interested in – the BRC approach is more adequate. After all, it focuses on expected future success, and not on past performance. H_0 is not accepted because it is considered likely to be true, but because it is sufficiently likely to be *predictively successful.*

Frequentists may object that Bernardo's approach is a very complicated way to obtain a simple result. After all, if we use *confidence intervals* instead of p-values, we will be able to appreciate the small effect size as well as the fact that the data are incompatible with the null hypothesis. A similar point can be made in Mayo's (1996) error-statistical framework: only a small discrepancy from the null hypothesis is warranted with a high degree of severity, but no discrepancy that points to a substantial extrasensory influence rather than to a tiny bias in the machine. Hence, Lindley's Paradox seems to vanish in thin air if we only adopt the right frequentist perspective.

To this point I have a twofold reply: First, confidence intervals and severity functions are, on a mathematical level, intimately connected to p-values and Neyman-Pearson error probabilities. Therefore they share a lot of the foundational problems of p-values, some of which have been mentioned above (see Royall 1997, for an elaborate discussion). A fully convincing reply to these criticisms is still pending. Second, confidence intervals do not involve a decision-theoretic component; they are interval estimators. They do not determine whether a precise null hypothesis should be accepted or rejected. (The case is a bit more complicated for Mayo's error statistics, but as I understand her, the kind of inferences she wants to make is about *severely warranted discrepancies from the null*, and not about decisions to accept or to reject a point null hypothesis.) If we take statistical tests to be serious decision problems, if the word "test" is more than a dummy for our preferred inference problem, then those frequentist techniques do not provide a convincing account of hypothesis testing.

5 Conclusions

We have demonstrated how Lindley's Paradox – the extreme divergence of Bayesian and frequentist inference in tests of a precise null hypothesis with large sample size – challenges the standard methods of both Bayesian and frequentist inference. Neither frequentist significance tests nor subjective Bayesian inference provides a convincing account of the problem. Therefore, I have introduced Bernardo's Bayesian Reference Criterion (BRC) as a full Bayesian, albeit not subjectivist model of testing a precise null hypothesis. It turns out that

BRC gives a sensible Bayesian treatment of Lindley's Paradox, with a focus on predictive performance and likely replication of the effect in deciding whether to accept or to reject the null. The motivation of BRC also exhibits a notable similarity to ideas voiced by Karl Popper.

Of course, Bernardo's reference Bayesian approach is not immune to objections. But anyway, BRC underlines that Bayesian inference in science need not necessarily infer to highly probable models – a misconception that is perpetuated in post-Carnapian primers on Bayesian inference and that has attracted Popper's understandable criticism. To provide some evidence: Howson and Urbach (1993, xvii) claim that "scientific reasoning is essentially reasoning in accordance with the formal principles of probability" and Earman (1992, 33) even takes, in his exposition of Bayesian reasoning, the liberty of announcing that "issues in Bayesian decision theory will be ignored". As argued in the paper, such a purely probabilistic Bayesianism falls short of an appropriate model of scientific reasoning.

In other words, Bayesianism should not be separated from its decision-theoretic component that involves, beside the well-known probabilistic representation of uncertainty, also a utility function of equal significance. Failure to appreciate this fact is, to my mind, partly responsible for the gap between the debates in statistical methodology and confirmation theory. This paper makes an attempt to bridge it.

References

Berger, J.O., and M. Delampady (1987): "Testing Precise Hypotheses", *Statistical Science* 2, 317–352.

Berger, J.O., and T. Sellke (1987): "Testing a point null hypothesis: The irreconciliability of P-values and evidence", *Journal of the American Statistical Association* 82, 112–139.

Bernardo, J.M. (1979): "Reference posterior distributions for Bayesian inference", *Journal of the Royal Statistical Society B* 41, 113–147.

Bernardo, J.M. (1999): "Nested Hypothesis Testing: The Bayesian Reference Criterion", in J. Bernardo et al. (eds.): *Bayesian Statistics 6: Proceedings of the Sixth Valencia Meeting*, 101–130. Oxford: Oxford University Press.

Bernardo, J.M. (2012): "Integrated objective Bayesian estimation and hypothesis testing", in J.M. Bernardo et al. (eds.): *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*, 1–68. Oxford: Oxford University Press.

Cohen, J. (1994): "The Earth is Round ($p < .05$)", *American Psychologist* 49, 997-1001.

Earman, J. (1992): *Bayes or Bust?*. Cambridge/MA: The MIT Press.

Fisher, R.A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.

Good, I.J. (1952): "Rational Decisions", *Journal of the Royal Statistical Society B* 14, 107–114.

Goodman, S.N. (1999): "Towards Evidence-Based Medical Statistics. 1: The P Value Fallacy", *Annals of Internal Medicine* 130, 1005–1013.

Howson, C. and P. Urbach (1993): *Scientific Reasoning: The Bayesian Approach*. Second Edition. La Salle: Open Court.

Jahn, R.G., B.J. Dunne and R.D. Nelson (1987): "Engineering anomalies research", *Journal of Scientific Exploration* 1, 21–50.

Jefferys, W.H. (1990): "Bayesian Analysis of Random Event Generator Data", *Journal of Scientific Exploration* 4, 153–169.

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

Mayo, D.G. (1996): *Error and the Growth of Experimental Knowledge*. Chicago & London: The University of Chicago Press.

Popper, K.R. (1934/59): *Logik der Forschung*. Berlin: Akademie Verlag. English translation as *The Logic of Scientific Discovery*. New York: Basic Books, 1959.

Popper, K.R. (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Harper.

Royall, R. (1997): *Scientific Evidence: A Likelihood Paradigm*. London: Chapman & Hall.