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

Jan Sprenger*

August 26, 2013

Abstract

Testing a point null hypothesis is a classical, but controversial issue in statistical methodology. A prominent illustration is Lindley's Paradox which emerges in hypothesis tests with large sample size and exposes a salient divergence between Bayesian and frequentist inference. A close analysis of the paradox reveals that both Bayesians and frequentists fail to satisfactorily resolve it. As an alternative, I suggest Bernardo's (1999) 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; (iii) it convincingly addresses 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. The author wishes to thank the Netherlands Organisation for Scientific Research (NWO) for support of his research through Veni grant 016.104.079, as well as José Bernardo, Cecilia Nardini and the audience at PSA 2012, San Diego, for providing helpful input and feedback

1. Introduction. Lindley's Paradox.

Lindley's Paradox exposes a salient divergence between subjective Bayesian and frequentist reasoning when a parametric point null hypothesis $H_0 : \theta = \theta_0$ is tested against an unspecified alternative $H_1 : \theta \neq \theta_0$. Since the paradox has repercussions for the interpretation of statistical tests in general, it is of high philosophical interest.

To illustrate the paradox, we give an example from parapsychological research (Jahn, Dunne and Nelson 1987). The case at hand involved the test of a subject's claim to affect a series of randomly generated zeros and ones ($\theta_0 = 0.5$) by means of extrasensory capacities (ESP). The subject claimed that his ESP would make the sample mean differ significantly from 0.5.

A very large dataset (N = 104.490.000) was collected to test this hypothesis. 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 $H_1 : \theta \neq 1/2$.

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

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

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

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

Thus, the data would be interpreted as strong evidence for the presence of 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, a uniform prior over the alternative, and calculates a Bayesian measure of evidence in favor of the null, namely the *Bayes factor*. 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:

$$B_{01}(x) := \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} \approx 12$$

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

This divergence between Bayesians and frequentists has, since the seminal paper of Lindley (1957), been known as *Lindley's Paradox*. In Lindley's original

formulation, the paradox is stated as follows: Assume that we compare 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). In that case, as N increases, the Bayesian evaluation of the data will become ever more inclined toward the null hypothesis. Thus, a result that seems to refute the null from a frequentist point of view can strongly support it from a Bayesian perspective. Put formally (for the case of Gaussian models):

> **Lindley's Paradox:** In a Gaussian 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 $1 - \alpha$. (cf. Lindley 1957, 187)

As the ESP example makes clear, Lindley's Paradox actually extends beyond Gaussian models with known variance. It exposes a general divergence between Bayesians and frequentists in hypothesis tests with large sample size.

In this paper, I consider the following questions: First, which statistical analysis of the ESP example, and Lindley's Paradox in general, is most adequate? Second, which implications does Lindley's Paradox have for the methodological debates between Bayesians and frequentists? Third, does our analysis have ramifications for the interpretation of point null hypothesis tests? I will argue that both the subjective 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 approach (the Bayesian Reference Criterion or BRC) holds considerable promise as a decision model of hypothesis testing, both in terms of the implied utility structure and as a reply to Lindley's Paradox.

2. Testing a precise null hypothesis: 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 be safely assumed that even the most unassuming placebo will have some minimal effect, positive or negative.

The answer is that precise null hypotheses often give us a useful idealization of reality. This is 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 the postulated mean value θ_0 of a Normal model, one canonically uses the 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 takes a value at least as high as $z(x)$ would be very small if the null were indeed true. Conventionally, $p < 0.05$ means significant evidence against the null and $p < 0.01$ very significant evidence. In the context of hypothesis testing, it is then common to say that the null hypothesis is rejected at the 0.05 level, etc.

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 $p(H_0)$. Bayes' Theorem can be used to calculate the posterior on the basis of the prior and the likelihood of the data.

If we investigate the source of Lindley's Paradox, one might conjecture that an "impartial", but unrealistically high prior for H_0 (e.g., $p(H_0) = 1/2$) is the culprit for the high posterior probability of the null. However, Lindley's findings persist if the analysis is conducted in terms of Bayes factors, like in the ESP example. These measures of evidence are independent of the particular prior of H_0. For instance, if the prior over the 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

$$\begin{aligned} B_{01}(x) &= \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{p(x|H_0)}{p(x|H_1)} \\ &= \sqrt{1 + \frac{N\tilde{\sigma}^2}{\sigma^2}}\, e^{\frac{-Nz(x)^2}{2N+2\sigma^2/\tilde{\sigma}^2}}, \end{aligned}$$

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.

How come that this result diverges so remarkably from the frequentist finding of significant evidence against the null? 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 remote from θ_0. By contrast, significance tests do not consider the likelihood of the data under these alternatives.

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 that favor 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 $\bar{X} \to \theta_0$, an ever greater chunk of the alternative H_1 will be far away from $\bar{X}$, favoring the null hypothesis.

These phenomena exemplify more general and foundational criticisms of frequentist inference, in particular the objection that p-values grossly overstate evidence against the null (Cohen 1994; Royall 1997; Goodman 1999). For instance, even the *minimum* of $p(H_0|x)$ under a large class of priors is typically much higher than the observed p-value (Berger and Sellke 1987).

Still, also the subjective Bayesian stance on hypothesis tests is not entirely satisfactory. Assigning a strictly positive degree of belief $p(H_0) > 0$ to a precise 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 (Lebesgue-)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 hypothesis $\theta = \theta_0$. Then, it would make sense to use strictly positive priors. Indeed, it has been shown that point null hypothesis tests approximate, in terms of Bayes factors, 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 $N \to \infty$, the convergence results break down, and testing a point null is no more analogous to testing whether a narrow interval contains θ (Bernardo 1999, 102). 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.

This may be the toughest challenge posed by Lindley's Paradox. In the debate with frequentists, Bayesians like to appeal to "foundations", but assigning a strictly positive probability to a precise hypothesis is hard to justify as a foundationally sound representation of subjective uncertainty.

Moreover, the Bayesian analysis fails to explain why hypothesis tests have such an appeal to scientific practitioners, even to those that are statistically well educated. Why should we bother testing a hypothesis if only posterior probabilities are relevant? Why even consider a precise hypothesis if it is known to be wrong? The next section will highlight these questions and briefly discuss the function of hypothesis tests in scientific inquiry.

3. Intermezzo: A note on precise hypotheses.

Since both Bayesians and frequentists struggle to deliver satisfactory responses to Lindley's Paradox, one may conjecture that the real problem is with testing a precise hypothesis as such. For instance, if we constructed a 95% confidence interval for θ in the ESP case, it would not include θ_0. But on the other hand, it would be close enough to θ_0 as to avoid the impression that the null was grossly mistaken.[1] Hence, Lindley's Paradox seems to vanish in thin air if we only adopt

[1] A similar point can be made in the error-statistical framework (Mayo 1996): only a small discrepancy from the null hypothesis would be warranted with a high degree of severity. Mayo

a different frequentist perspective.

However, this proposal is not satisfactory either. Confidence intervals do not state which hypotheses are *credible* – they only list the hypotheses that are *consistent* with the data, in the sense that these hypotheses would not be rejected in a significance test. Therefore, confidence intervals are intimately connected to significance tests and share a lot of their foundational problems (cf. Seidenfeld 1981; Sprenger 2013). Second, confidence intervals do not involve a decision-theoretic component; they are interval estimators. In particular, they do not explain why tests of a precise null have any role in scientific methodology. Since any proper resolution of Lindley's Paradox should address this question, a confidence interval approach evades rather than solves the paradox.

On this note, one ought to realize that tests of a precise null usually serve two purposes: to find out whether an intervention has a significant effect, and, since any intervention will have *some* minute effect, to decide whether the null hypothesis can be used as a *proxy* for the more general model. The point null is usually much easier to test and to handle than any composite hypothesis, so we have positive reasons to "accept" it, as long as the divergence to the data is not too large.

This view of scientific inference is hardly compatible with an orthodox Bayesian approach. For instance, the assumption $p(H_0) > 0$ neglects that hypothesis tests ask, in the first place, if H_0 is a reasonable simplification of a more general model—and not if we entertain a high degree of belief in a precise value of θ. Also, point null hypothesis tests are by definition asymmetric, but a subjective Bayesian analysis in terms of Bayes factors or posterior probabilities is essentially symmetric.

In total, subjective Bayesians have a hard time to explain why informative and precise, but improbable hypotheses should sometimes be preferred over more general alternatives. The challenge for the Bayesian consists in modeling that we may be less interested in the *truth* of H_0 than in its *usefulness*. The next section presents an answer to this effect developed by José Miguel Bernardo (1999, 2012).

4. The BRC approach to hypothesis testing.

This section presents a proposal for full Bayesian decision model for point null hypothesis testing that addresses Lindley's Paradox: José Bernardo's Bayesian Reference Criterion or BRC (Bernardo 1999, 2012). The point consists in shifting the focus from the truth of H_0 to its *predictive value* and in stipulating a specific utility structure. While classical Bayesian accounts of hypothesis testing involve simple exogenous utilities (e.g., a loss of zero for correct decisions, and one for

speaks about acceptances and rejections, too, but in fact, she is interested in *severely warranted discrepancies from the null*, not in decisions to accept or to reject a point null hypothesis.

wrong decisions) and use the posterior probability as the only criterion for accepting or rejecting the null, Bernardo's approach is based on endogenous, prediction-based utilities. In the remainder, I sketch a simplified version of Bernardo's BRC in order to elaborate the main ideas of philosophical interest.

Since the work of R. A. Fisher, the *replication* of previously observed effects has been recognized as a main goal of experimental research in science and as a main motivations for significance tests (cf. Schmidt and al. 1997). Therefore, a central component of Bernardo's decision model focuses on the expected predictive accuracy of the null for future data. 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 scoring rule 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

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

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. As explained above, H_0 should be preferred to H_1 ceteris paribus 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 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. Much better than subjective Bayesianism, this approach represents 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

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

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" $d^*(\alpha, c_0 - c_1)$. For the latter, Bernardo suggests a conventional choice that recovers the well-probed scientific practice of regarding five standard deviations as strong evidence against the null.[2] 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, Bernardo 1999):** 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
>
> $$\int_{\theta\in\Theta} p(\theta|x) \left(\int \log \frac{p(y|\theta)}{p(y|\theta_0)}\, p(y|\theta)\, dy \right) d\theta > d^*(\alpha, c_0 - c_1). \qquad (1)$$

This approach has a variety of remarkable features. First, it puts hypothesis testing on firm decision-theoretic grounds with predictive value being the primary criterion. This foundational soundness distinguishes BRC vis-à-vis frequentist procedures.

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

[2]This evidential standard was also used in the recent discovery of the Higgs particle. For Bayesian justifications of this practice, see Berger and Delampady (1987) and Berger and Sellke (1987).

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 notable 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). In marked difference to the orthodox Bayesian appraoch, accepting H_0 does no more involve commitment to the truth or likelihood of H_0.

Third, the approach is better equipped than subjective Bayesianism to account for frequentist intuitions, since under some conditions, the results of BRC 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.

5. 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). In other words, the predictive loss from taking the null as a proxy for the posterior-corrected alternative will be substantial.

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. In this sense, we are justified to reject the null without having to infer to a substantial discrepancy between θ and θ_0.

Although BRC has a sound basis in Bayesian decision theory, the results of a BRC analysis disagree with Jeffery's subjective Bayesian analysis. Why is this the case? First, the conventional utility structure is substantially changed in BRC, and the final decision is no more a simple function of the posterior probability of H_0. 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. Thus, BRC indeed gives a convincing account of Lindley's Paradox in the ESP example.

6. Conclusion and Outlook.

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 a classical frequentist nor a subjective Bayesian analysis provide a convincing account of the problem. Therefore, I have presented Bernardo's Bayesian Reference Criterion (BRC) as a full Bayesian model of testing point null hypotheses. It turns out that BRC gives a sensible Bayesian treatment of Lindley's Paradox, due to its focus on predictive performance and likely replication of the effect. Although BRC has sound foundations in Subjective Expected Utility Theory, it preserves testing a precise hypothesis as a distinct form of statistical inference and can be motivated from a broadly Popperian perspective.

Of course, the BRC approach is not immune to objections (see the discussion pieces in Bernardo 2012). However, BRC definitely 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 understandable criticism. For instance, Earman (1992, 33) takes, in his exposition of Bayesian reasoning, the liberty of announcing that "issues in Bayesian decision theory will be ignored". Contrary to Earman, I claim that Bayesian reasoning cannot dispense with the decision-theoretic dimension if it aims at scientific relevance. A purely epistemic approach to theory choice, as exemplified in much of Bayesian confirmation theory, falls short of an appropriate model of scientific reasoning.

Therefore, this paper is not only a contribution to statistical methodology: it highlights the need to appreciate the subtle interplay of probabilities and (predictive) utilities in Bayesian inference, and to change our perspective on the use of Bayesian reasoning in science.

References

Berger, James O., and Mohan Delampady. 1987. "Testing Precise Hypotheses." *Statistical Science* 2:317–352.

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

Schmidt, Frank L., and John E. Hunter. 1997. "Eight Common but False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data." In *What if there were no significance tests?*, ed. Lisa L. Harlow et al., 37–64. Mahwah/NJ: Erlbaum.

Seidenfeld, Teddy. 1981. "On after-trial properties of best Neyman-Pearson confidence intervals." *Philosophy of Science* 48:281–291.

Sprenger, Jan. 2013. "Bayesianism vs. Frequentism in Statistical Inference". Forthcoming in *Oxford Handbook of Probability and Philosophy*, ed. A. Hájek and C. Hitchcock, Oxford: Oxford University Press.