

# How Do Hypothesis Tests Provide Scientific Evidence? Reconciling Karl Popper and Thomas Bayes

June 17, 2011

## **Abstract**

Testing a null hypothesis is the most common activity in statistical data analysis. Yet, the proper interpretation of such tests is still controversial. I argue that this is not only a central question of philosophy of statistics, but also at the heart of the disagreement between Popperians and Bayesians about scientific method. After elaborating on the shortcomings of the prevalent interpretations, including the Fisherian and the subjective Bayesian approach, I propose to regard frequentist hypothesis tests as stylized Bayesian decision problems, thereby solving refractory problems such as Lindley's paradox. The crucial tool is the reference Bayesian approach, developed by the statistician José Bernardo. By exploring its philosophical implications, the paper provides (i) a foundationally sound account of hypothesis testing, (ii) a way to explicate Popper's account of empirical corroboration, and (iii) a partial reconciliation of Popperian and Bayesian perspectives on scientific inference.

# 1 Introduction. Hypothesis tests, corroboration and Bayesianism

The title question of this article – how do hypothesis tests provide scientific evidence? – seems to be silly. The practice of science apparently teaches us the answer. Einstein’s General Theory of Relativity was famously tested by Eddington’s observation of the 1919 solar eclipse. Econometricians test models of time series for autoregression, heteroscedasticity, or moving averages. Psychologists use significance tests to infer causal effects. And so on.

The statistical foundations and epistemic justifications of hypothesis testing are, however, controversial. Strictly speaking, the outcome of a test is a decision to either accept or to reject a hypothesis. Indeed, statistical hypothesis tests in industrial quality control, pharmaceutical research or even the courtroom serve the purpose of reaching a decision on the basis of statistical data, and taking action accordingly. Such a decision could mean to deny or to accept a delivery of goods that has been sampled for defect elements, to stop or to continue the development of a medical drug, to sentence or to acquit a defendant. Hypothesis tests become rules of “inductive behavior” (Neyman and Pearson 1933) where an evidential assessment of the tested hypotheses is not of intrinsic interest, but just in so far as it helps us taking the right action.

In such practical applications, we take into account the possible consequences of our actions, and design our tests accordingly. For instance, in legal trials it is more important not to convict an innocent than to acquit a culprit. However, since the early days of statistical reasoning, it has been argued that such considerations do not suit the epistemic goals of *science*:

In the field of pure research no assessment of the cost of wrong conclusions [...] can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence. (Fisher 1935, 25–26)

Two arguments are implied in this quote of a famous statistician, methodologist and scientist. First, we cannot quantify the utility that correctly accepting or rejecting a hypothesis will eventually have for the advancement of science. The far-reaching consequences of such a decision are beyond our horizon. Second, statistical hypothesis tests should state the *evidence* for or against the tested hypothesis: a scientist is interested in empirical support for the empirical adequacy or inadequacy of a theory. Such a judgment should not be obscured by the practical long-term consequences of working with this rather than that hypothesis. Both arguments push us to explicate how hypothesis tests can support scientific inference, instead of just calibrating our *behavior* with statistical data.

However, due to the widespread disagreements about foundations of statistical inference, there is no consensus on this question, neither in statistics, nor in the sciences, nor in philosophy.

Of course, this question is crucial for the methodology of statistical inference. But it is also of vital significance for general philosophy of science, and in particular for Karl R. Popper's critical rationalism. According to Popper, science proceeds by severely testing hypotheses, that is, by sincerely attempting to prove them wrong. While Popper remains skeptical that the survival of severe tests ever justifies belief in the *truth* of a hypothesis, he realizes that scientists often have to form preferences over theories in order to decide which predictions should be relied on, which theory should be taken as a basis for further development, and which ones should be dismissed:<sup>1</sup>

there is not even anything irrational in relying for practical purposes upon well-tested theories, for no more rational course of action is open to us (Popper 1963, 51)

Independent of whether "action" refers to prediction or theoretical development: this attitude requires a positive account of testing and (tentative) theory choice. Admittedly, Popper's writings on what sort of rational preferences are pragmatically justified are not unambiguous, and sometimes have a skeptical twist (e.g., Popper 1956/83, 64–65). My take on Popper tries to resolve cases of doubt in a way that Popper's thoughts can be applied to central discussions in statistical methodology.

Popper calls the degree to which a hypothesis has been severely and successfully tested the *corroboration* of a hypothesis, which may also be interpreted as a "measure of the rationality of our beliefs" in a scientific hypothesis (Popper 1934/59, 414). For being in sync with accepted scientific practice – a point that Popper has repeatedly claimed in favor of his own approach, and against Carnap's (1950) inductivism – Popper needs to explicate the notions of corroboration, falsification and severity in the context of statistical hypothesis tests. Given the explosion of statistical reasoning in the sciences, the critical rationalist cannot possibly want to restrict her methodology to the testing of deterministic hypotheses alone.

Thus, a critical rationalist like Popper needs to explain what corroboration means in statistical hypothesis tests. It is well known that he eschewed the criteria of high posterior probability or incremental Bayesian confirmation because such hypotheses would neither be informative, predictively interesting nor ex-

---

<sup>1</sup>See also the following quote from "Objective Knowledge" (Popper 1972/79, 21-22): "From a rational point of view we should not rely on any theory, for no theory has been shown to be true, or can be shown to be true [...] But we should prefer as basis for action the best-tested theory."

planatory valuable. “As scientists, we do not seek highly probable theories, but explanations; that is to say, powerful and improbable theories.” (Popper 1963, 58) This skeptical attitude towards probabilistic explications conflicts, however, with the measure of degree of corroboration that Popper developed himself in his 1954 *BJPS* paper (reprinted in Popper 1934/59, 401):

$$c(H, E) = (1 + p(H)p(H|E)) \frac{p(E|H) - p(E)}{p(E|H) + p(E)} \quad (1)$$

First,  $c(H, E)$  is an increasing function of the posterior probability of  $H$ ,  $p(H|E)$ . Second, to calculate  $c(H|E)$ , one would have to know the prior probability of  $h$ . This conflicts with the Popperian scepticism towards assignments of non-zero prior probabilities to sufficiently general hypotheses.<sup>2</sup> While Popper’s measure may be a valuable contribution to Bayesian confirmation theory (e.g., Crupi, Tentori and Gonzalez 2007), it does not fit Popper’s general methodological approach, and I will neglect it in the remainder.

So far I have identified two goals of inquiry: an evidential interpretation of statistical hypothesis tests, and an explication of Popperian corroboration. A third line of inquiry concerns Popper’s opposition to Bayesianism, the dominant account of inductive learning in philosophy of science. The basic principles of Bayesianism state that agents entertain subjective degrees of belief in a hypothesis, that these degrees of belief can be represented by a probability function  $p(\cdot)$ , and that we learn from experience by means of conditionalizing on evidence  $E$ :

$$p_{\text{new}}(H) := p(H|E) = p(H) \frac{p(E|H)}{p(E)}. \quad (2)$$

The inferences that we make, and the posterior degrees of belief that we have, qualify as rational because they emerge as the result of a rational belief revision process. Now, the question at stake is not whether Bayesian Conditionalization is a valid belief revision rule. This is usually accepted as uncontentious. The question is rather whether Bayesianism is a viable model of *scientific* rationality. Proponents claim that “scientific reasoning is essentially reasoning in accordance with the formal principles of probability” (Howson and Urbach 1993, xvii). The Bayesian intuition is that by gathering evidence and updating our beliefs, we find out which hypotheses are best supported and therefore most likely to explain the phenomena which we are studying. Accordingly, high posterior probability becomes a measure of the acceptability of a hypothesis, and scientific inference is based on this posterior distribution of beliefs. This variety of Bayesianism will be referred to as subjective Bayesianism, in opposition to approaches that build on objective priors. Arguably, it is the most common account of Bayesianism in philosophy of science (Earman (1992, 142); Joyce (1998); Talbott (2008)).

---

<sup>2</sup>See Rowbottom (2011) for a more extended criticism of this measure.

Apparently, there is an unsurmountable gap between Popperians and Bayesians. The first reason is Popper’s dismissal of the subjective interpretation of probability in science (e.g., in Popper 1956/83). But from a methodological point of view, the second reason is more important. While Bayesians synthesize evidence and prior beliefs into a posterior probability distribution over the available hypotheses, Popper denies that such a distribution has any inferential value (for further discussion, see Rowbottom 2010).

This paper explores the merits of a particular statistical approach – José Bernardo’s (1999, 2011) *reference Bayesianism* – as a solution to the three problems described above. We do not only believe that Bernardo’s account is a foundationally sound (albeit perhaps not the only one) way of interpreting hypothesis tests: it also captures several of Popper’s methodological aims and bridges, as an extra benefit, the gap between Popperians and Bayesians. In other words, the paper essentially provides (i) a foundationally sound account of hypothesis testing, (ii) a way to explicate Popper’s account of corroboration, and (iii) a partial reconciliation of Popperian and Bayesian perspectives on scientific inference. Specifically, it is argued that Popper’s anti-Bayesianism is no natural consequence of his views on testing, falsification and corroboration.

To achieve these goals, I start out arguing that tests of precise, two-sided (null) hypotheses should be the focus of the inquiry (section 2). Second, I make use of Lindley’s paradox to criticize Fisher’s classical account of significance testing as well as the subjective Bayesian approach (section 3). Third, I argue that the *reference Bayesian* account of hypothesis testing (Bernardo 1999) fares much better in this respect: we are able to embed a Popperian take on the goal and methods of science into a genuine Bayesian model of hypothesis testing (section 4) and to explain the Bayesian-frequentist divergence in Lindley’s paradox (section 5).

## 2 Testing a Precise Null Hypothesis

For generating scientific knowledge, Popper famously proposed the method of conjecture and refutation, of trial and error. This implies in particular that the hypotheses to be tested need (and should) not be probable, but rather to the contrary: those hypotheses are distinguished by their testability, explanatory power and empirical content, virtues that are in tension with high probability. The greater the logical strength and empirical content of a hypothesis, the less probable it is, and vice versa. Only such hypothesis take a serious risk of being falsified, which is a hallmark of science vis-à-vis pseudo-science.

Thus, Popper concludes that “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 – like in statistical reasoning – not understood as endorsing the theory’s truth, but as providing guidance for predictions and further theoretical developments. We should not “rely” on it in any strong sense, but only tentatively, subject to further testing.

This focus on testing powerful and improbable theories strongly suggests *tests of a precise null hypothesis*, that is, specific values of a parameter of interest, as the focus of our investigation. We may posit a particular value for Newton’s gravitational constant  $\gamma$ , but the alternative, that the value of  $\gamma$  is different, is not of comparable precision and simplicity. The same would hold for tests of causal independence between two variables: an empirical refutation of independence is, in itself, no proof of a particular degree of dependence – for that, new tests would be required. Scientific hypothesis testing is usually *asymmetric*: failure of a hypothesis to survive a severe test does not imply that the alternatives (or a particular hypothesis therein) did survive a severe test. Consequently, a “rejection” of the null should not be understood literally – remember that we are not interested in a behavioristic interpretation –, but as stating strong evidence against the null, as an incompatibility statement between data and hypothesis.

The simplest instance of this testing problem is a statistical model with an unknown parameter of interest  $\theta \in \mathbb{R}^k$ . We would like to test the null hypothesis  $H_0$  that  $\theta$  is equal to a specific value  $\theta = \theta_0$ . Hypotheses of that form are as informative and risky as it gets. As suggested above, that value  $\theta_0$  could symbolize a natural constant, stand for causal independence between two variables, etc. This precise null hypothesis is then tested against the unspecified alternative  $H_1 : \theta \neq \theta_0$  standing for any other value of  $\theta$ , within the prespecified statistical model. This *two-sided testing problem* will be our battleground for developing a valid evidential interpretation of hypothesis tests and for fleshing out Popperian corroboration.<sup>3</sup>

It could be objected that a severe test of a hypothesis should test this hypothesis *in isolation*, and not against representatives of the same statistical model. Gillies (1971, 245) has even called such alternatives “trivial variants of the original hypothesis”. However, it has been argued convincingly (e.g., Spielman 1974; Royall 1997) that the very idea of testing a precise hypothesis in isolation, without specifying a contrast class, is highly problematic and should be abandoned.

While two-sided hypothesis tests are a natural implementation of propos-

---

<sup>3</sup>To support this choice, note that two-sided tests are already a salient issue in statistical methodology (Berger and Delampady 1987; Berger and Sellke 1987) and arguably more general than Albert’s (1992) proposal to restrict Popperian falsification and corroboration to testing hypotheses where we measure results with finite precision only.

ing bold theories and subjecting them to severe empirical scrutiny, things are different for the similarly popular *one-sided tests*. The one-sided test of  $\theta$  partitions the parameter space into two subsets, e.g., it compares the hypothesis  $H_0 : \theta \leq \theta_0$  to the alternative  $H_1 : \theta > \theta_0$ . This might be a suitable scenario for testing whether the proportion of defect sundries, the effect of a medical drug, or the population size of an endangered species exceed a certain threshold  $\theta_0$ . We infer whether the true value of  $\theta$  is larger or smaller than  $\theta_0$ , but it is not a severe test in the sense that the tested hypothesis would be bold or particularly informative. Rather it is an essentially symmetric comparison of two hypotheses about  $\theta$ .

In particular, the error-statistical approach of Mayo (1996) and Mayo and Spanos (2006) focuses on the proper interpretation of these one-sided tests, and not on (two-sided) tests of precise nulls. Mayo (personal communication) is explicit that an error statistician *never* accepts a point null hypothesis, but only infers to a hypothesis that the discrepancy from the null value is greater or lesser than a specific value. Such hypotheses have the form  $\theta > \theta_0 + c$  or  $\theta \leq \theta_0 - c$ , with  $c > 0$ . Therefore, Mayo and Spanos do not address the problem of inferring or rejecting a precise hypothesis that corresponds to a useful scientific idealization (though see Mayo and Spanos 2004).

Therefore, both the behavioristic approach of Neyman and Pearson and the error-statistical developments of Mayo and Spanos do not fit the purpose of our inquiry. The remaining frequentist school is Fisherian significance testing: the evidence speaks strongly against the null when a discrepancy measure that is based on the null's probability density – the famous p-value – exceeds a specific value. We will discuss this approach and compare it to subjective Bayesian inference in one of the most refractory problems of statistical inference: the Jeffreys-Lindley effect, or in short, Lindley's paradox.

### 3 Significance Testing, Bayesian Inference and Lindley's Paradox

We have concluded the previous section by the observation that the kind of statistical hypothesis tests that are relevant for our purpose are two-sided tests of a precise null hypothesis. Throughout the paper, I will focus on the simple, canonical example of inferring the mean  $\theta$  of a Normally distributed random variable  $X \sim N(\theta, \sigma^2)$  with known variance  $\sigma^2$ . Despite its apparent simplicity and mathematical well-behavedness, it is sufficient to expose the fundamental conceptual differences between the various statistical schools of inference.

In this example, the null hypothesis  $H_0 : \theta = \theta_0$  asserts that the parameter of interest takes a precise value  $\theta_0$ , with an unspecified alternative  $H_1 : \theta \neq \theta_0$

representing the more general model. We also assume, for the sake of concentrating on foundational issues, that the overall probability model has not been grossly misspecified. A typical interpretation of such a null would be that a treatment effect is absent, that two variables are causally independent, that two populations do not differ in an important characteristic, etc.

Due to the influential heritage of R. A. Fisher, two-sided tests of such a null hypothesis are often conducted as *significance tests*. That is, they aim at detecting the presence of “significant effects”, e.g., when comparing a treatment group with a control group in a randomized clinical trial. 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  $x := (x_1, \dots, x_N)$  with respect to postulated mean value  $\theta_0$  of a Normal model, one canonically uses the standardized statistic

$$z(x) := \frac{\sqrt{N}}{\sigma} \left( \frac{1}{N} \sum_{i=1}^N x_i - \theta_0 \right). \quad (3)$$

Noting that  $z$  is essentially a function of the difference between sample mean and hypothesized parameter value, we can interpret higher values of  $z$  as denoting a higher divergence from the null, and vice versa.<sup>4</sup>

Since the distribution of  $z$  varies for most statistical models with the sample size, some kind of standardization is required. Practitioners usually use the *p-value* or *significance level*, that is, the “tail area” of the null under the observed data. This can be computed, in the case of a Normal model, as

$$p := P(|z(X)| \geq |z(x)|) = 2(1 - \Phi(|z(x)|)) \quad (4)$$

with  $\Phi$  denoting the cumulative distribution function of  $N(\theta_0, \sigma^2)$ . In figure 1, the critical region corresponds to the shaded area, that is, the tails of the Normal distribution. 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. Therefore, Fisher has interpreted p-values as “a measure of the rational grounds for the *disbelief* [in the null hypothesis] it augments” (Fisher 1956, 43).

---

<sup>4</sup> $z(x)$  has the important property of being a *sufficient statistic*, meaning that there is no function of the data that provides additional information about the parameter of interest. Mathematically, sufficiency is defined as independence between data and parameter, conditional on the value of the sufficient statistic:  $P(X = x | z(X) = y, \theta) = \Pr(X = x | z(X) = y)$ . Moreover,  $z$  is, for any distribution of the data, asymptotically distributed as  $N(0, 1)$ , if  $N \rightarrow \infty$ , due to the Central Limit Theorem, which explains why it is a preferred choice for statistical testing procedures.

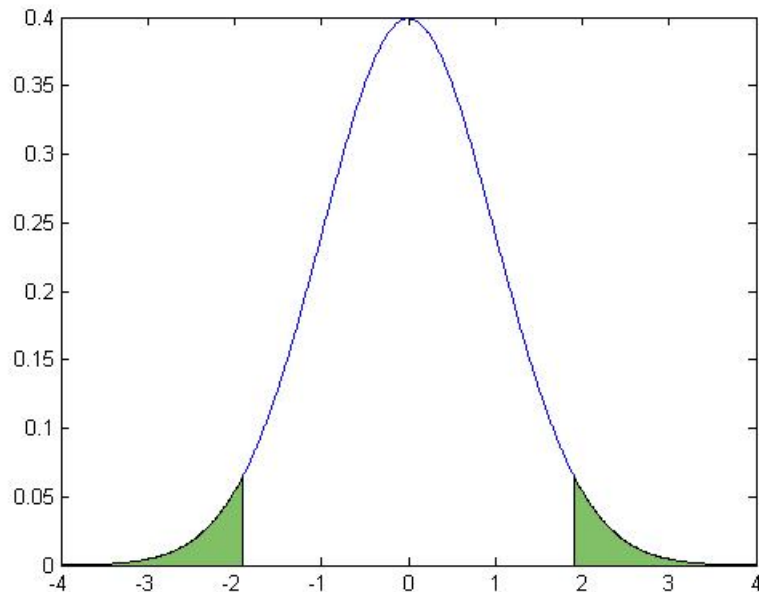


Figure 1: The critical region of the standard normal distribution  $N(0, 1)$  at the 0.05 level.

Subjective Bayesians choose a completely different approach to this inference problem. For them, scientific inference obeys the rules of probabilistic calculus. Prior probabilities represent honest, subjective degrees of belief, and are updated by means of Bayesian conditionalization. For a Bayesian, a null hypothesis has not withstood a severe test if its posterior probability  $p(H_0|E)$ , the synthesis of data  $E$  and prior  $p(H_0)$ , is sufficiently low.

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 Selke (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).

Admittedly, a staunch frequentist might deny the significance of a subjective Bayesian analysis, but in some cases (e.g., when reasoning in games of chance),

probability assignments to hypotheses under test can be objectively grounded. In these canonical cases, p-values should be in sync with a Bayesian analysis. So it seems that the significance tester has a real problem, even if we leave aside further technical objections and all the misinterpretations of p-values that frequently occur in practice (Cohen 1994; Harlow et al. 1997; Ziliak and McCloskey 2004).<sup>5</sup>

The case against significance testing is even stronger if we are dealing with high sample sizes. Consider the two-sided testing problem of  $X \sim N(\theta, \sigma^2)$ , with known variance  $\sigma^2$ , point null hypothesis  $H_0 : \theta = \theta_0$ , and the unspecified alternative  $H_1 : \theta \neq \theta_0$ . Assume further 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 sample size.

Perhaps, it comes as no surprise that the outcomes of a Bayesian and a frequentist analysis diverge for increasing  $N$ , as long as  $p(H_0) > 0$ . Arguably surprising, however, is the extent of that divergence: as  $N \rightarrow \infty$ , the posterior probability of the null,  $p(H_0|x)$  converges to 1 for almost any prior distribution over  $H_1$  (Lindley 1957)! A result that speaks highly significantly against the null from a frequentist point of view lends strong support to it from a Bayesian perspective. More precisely:

**Lindley’s Paradox:** Take a Normal model  $N(\theta, \sigma^2)$  with known variance  $\sigma^2$ ,  $H_0 : \theta = \theta_0$ ,  $H_1 : \theta \neq \theta_0$ , 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  $N(\alpha, p(\cdot))$  and independent, identically distributed (i.i.d.) data  $x = (x_1, \dots, x_N)$  such that

1. The sample mean  $\bar{x}$  is significantly different from  $\theta_0$  at level  $\alpha$ ;
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))

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:

$$B_{01}(x) := \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{p(x|H_0)}{p(x|H_1)}, \quad (5)$$

---

<sup>5</sup>James O. Berger’s (2003) *conditional frequentist approach* is designed to mitigate these problems by conditioning hypothesis tests on the exact value of an observed statistic. The idea of Berger’s approach is to make Bayesian and frequentist analyses agree in practice (though not necessarily on methodology).

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

$$B_{01}(x) = \sqrt{1 + \frac{N\tilde{\sigma}^2}{\sigma^2}} e^{\frac{-Nz(x)^2}{2N+2\sigma^2/\tilde{\sigma}^2}}, \quad (6)$$

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 any underlying probability model, in distribution against the  $N(0, 1)$  distribution, thus,  $c\sigma \approx \sqrt{N}(\bar{X} - \theta_0)$ . This implies that as  $N \rightarrow \infty$ , we obtain  $\bar{X} \rightarrow \theta_0$ . In other words, the sample mean gets ever closer to  $\theta_0$ , favoring the null over the alternatives. For the deviance between the averaged, 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  $H_0$ .

Note that in this asymptotic argument, nothing hinges on the fact that  $X$  is Normally distributed – it is indicative of a general phenomenon and holds for other distributions, too.

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  $\theta_0$  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  $\theta_0$  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  $\bar{X} \rightarrow \theta_0$ , if there is a “significant effect” of constant size in the data, 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. The first objection is foundational: assigning a strictly positive prior probability, that is, degree of belief  $p(H_0) > 0$  to the point value  $\theta_0$  is a misleading and inaccurate representation of our subjective uncertainty. In terms of degrees of belief,  $\theta_0$  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  $\epsilon$ , 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 when evaluating a hypothesis test in terms of Bayes factors, testing a precise null approximates 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  $\theta$ . 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  $\theta_0$  – 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  $\theta$ .

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.

We might decide to make evidential judgments by means of Bayes factors, not by means of posterior probabilities. Still, it is not clear whether the analysis tracks the right target. An analysis in terms of Bayes factors effectively amounts to comparing the likelihood of the data under the null to the *averaged*, integrated likelihood of the data under the alternative. But is this a test of the null hypothesis against a more general model, a test for compatibility of the null with the data, rather than comparing it to a specific *mixture* of alternatives?

The bottom line of all this is that the essential symmetry of a subjective Bayesian analysis makes it inadequate for explicating the test of a precise null against an unspecified alternative, and a fortiori, for explicating Popperian corroboration. It fails to explain why hypothesis tests have such an appeal to

scientific practitioners, some of whom defend them although they are aware of the problems of assigning them a valid evidential interpretation.<sup>6</sup> The very idea of testing a hypothesis is in principle alien to Bayesianism. Responding by “so much the worse for the non-Bayesian” displays, to my mind, an overly narrow Bayesians-in-the-sky attitude. First, hypothesis tests are an integral and well-entrenched part of scientific research. Instead of discarding them out of hand, it might be more interesting to develop a Bayesian perspective where they need not be regarded as an evil. This is the project that I pursue in the rest of the paper. Second, a pure Bayesian analysis is often impractical because of the difficulties of eliciting meaningful subjective prior distributions, and the associated computational problems. Similarly, the Bayesian has a hard time to explain why informative, predictively interesting hypotheses should sometimes be preferred over more general, but less committing and less interesting alternatives.

There is another problem, too. Scientists often shun Bayesian measures of evidence because of their “subjectivity” that clashes with an unattainable, but perhaps useful ideal of scientific objectivity. In particular, since there are no hard restrictions on how skewed the prior distribution over the hypotheses in  $H_1 : \theta \neq \theta_0$  may be, it remains unclear why we should accept a certain posterior probability, a certain Bayes factor as an authoritative quantification of the evidence. An extreme opinion is as good as any other. P-values seem, despite all shortcomings, to avoid this problem.

This last problem suggests a resolution in terms of *default prior probabilities*: priors that depend, to some extent, on the design and the sample size of the experiment such that they are responsive to the expected strength of the evidence. The next section explores, building on José Bernardo’s reference Bayesianism, how Bayesian and frequentist intuitions can be brought together along these lines.

## 4 Expected Information as Expected Utility

The main idea of Bernardo’s proposal, which is the subject matter of this section, is to understand a hypothesis test as a proper decision problem where we make a decision on whether or not to 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. Instead of favoring the highly probable hypotheses, Bernardo’s suggestion incorporates the virtues of testability, informativity, corroborability into a decision to reject or to accept the null. This is in line with Popper’s view that “science does not aim, primarily, at high

---

<sup>6</sup>This phenomenon is manifested in, e.g., the contributions in Harlow et al. (1997).

probabilities. It aims at high informative content, well backed by experience.” (Popper 1934/59, 399).

To be able to accommodate these Popperian intuitions, 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 we know as subjective or probabilistic Bayesian inference:

the more traditional Bayesian approaches to model comparison, such as determining the posterior probabilities of competing models or computing the relevant Bayes factors, can be obtained as particular cases [...] by using appropriately chosen, stylised utility functions (Bernardo and Smith 1994, 420)

In the remainder, I simplify Bernardo’s (1999, section 2-3) account 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 utility function is expected predictive accuracy, and we need a function that evaluates the predictive score of a distribution, given some data  $y$ . The canonical approach for these purposes is the logarithmic score  $\log p(y|\theta)$  (Good 1952; Bernardo 1979b): 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. Using the natural logarithm has an additional advantage: if the data  $y$  consist of several independent observations  $(y_1, \dots, y_N)$ , then the overall predictive score amounts to the sum of the score for the single observations. This is warranted by the equality  $\log p(y_1, \dots, y_N|\theta) = \sum_i \log p(y_i|\theta)$ .

A generalization of this utility function describes the score of data  $y$  under parameter value  $\theta$  as  $q(\theta, y) = \alpha \log p(y|\theta) + \beta(y)$ , where  $\alpha$  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  $\theta$  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  $H_j$ .  $H_0$  is better testable and simpler to handle than  $H_1$  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  $\theta$ , and that the alternative is not represented by a probabilistic average (e.g., the posterior mean), but by its best element, namely  $\theta$ . This is arguably more faithful than subjective Bayesianism to the asymmetry of scientific hypothesis testing, and to the idea of severely testing a hypothesis against the unknown truth. Consequently, the difference in *expected utility*, conditional on the posterior density of  $\theta$ , can be written as

$$\int_{\theta \in \Theta} (U(H_1, \theta) - U(H_0, \theta)) p(\theta|x) d\theta$$

$$= \alpha \int_{\theta \in \Theta} \left( \int \log \frac{p(y|\theta)}{p(y|\theta_0)} p(y|\theta) \right) p(\theta|x) dy d\theta + \int \beta(y) p(y|\theta) dy - \int \beta(y) p(y|\theta) dy + c_0 - c_1$$

$$= \alpha \int_{\theta \in \Theta} \left( \int \log \frac{p(y|\theta)}{p(y|\theta_0)} p(y|\theta) dy \right) p(\theta|x) d\theta + c_0 - c_1.$$

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 scientific practice of regarding three standard deviations as significant 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 model  $(p(\cdot|\theta), \theta \in \Theta)$ , if and only if

$$\int_{\theta \in \Theta} \left( \int \log \frac{p(y|\theta)}{p(y|\theta_0)} p(y|\theta) dy \right) p(\theta|x) d\theta > d^*(c_0 - c_1). \quad (7)$$

This approach has a variety of remarkable features. First, it puts hypothesis testing on firm decision-theoretic grounds. The decision to accept  $H_0$  is a rational decision given the particular goals one has set. Second, accepting the null,

that is, using  $\theta_0$  as a proxy for  $\theta$ , amounts to claiming that the difference in the expected predictive success of  $\theta_0$  and the true parameter value  $\theta$  will be offset by the greater testability, informativity and simplicity of  $H_0$ . This is precisely what Popper referred to in the quote at the beginning of this section: a hypothesis survives a severe test if it has “high informative content, well backed by experience”. 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  $\theta$  or a 1:1-transformation  $g(\theta)$ .

Obviously, BRC depends on the posterior probability function  $p(\theta|x)$ , and thus on the priors implied. In principle, it is neutral as to whether they are generated in a subjective way, or by some conventional procedure. The BRC model of hypothesis testing therefore stands open to a subjective Bayesian, too. In practice, however, we will often face the problem that reliable subjective information is hard to get, or too costly to elicit. For that case, Bernardo suggests the use of *reference priors* as a default – priors that maximize, relative to a chosen experimental design, the information in the data.<sup>7</sup> In other words, reference priors specify what the result of an experiment would be if we chose priors as to maximize the information which the experiment conveys (Bernardo 2011, 60).<sup>8</sup>

The use of reference priors in the context of hypothesis testing is motivated by the observation that, whenever we have a comparably few data, and only vague knowledge of the values of the parameter  $\theta$ , it certainly makes sense to be cautious with respect to prior information. Vice versa, if we know the data set to be huge, we can allow for a more skewed prior because we will probably observe strong evidence for either hypothesis. Therefore these priors they can be defended as a useful reference or default approach. Let us now see how they can be applied to Lindley’s paradox.

## 5 Revisiting Lindley’s Paradox

Lindley’s paradox is arguably the most famous illustration of how (subjective) Bayesian and frequentist approaches to null hypothesis testing can fall apart. In this section, we deal with Binomial data from a real study. The case at hand involves a test of the claim of a subject to possess extrasensory capacities,

---

<sup>7</sup>That notion is explicated by maximizing  $\int_{\Theta} \int p(x)p(\theta) \log \frac{p(x)p(\theta)}{p(x,\theta)} dx d\theta$ , the expected association between the data and the parameter of interest, for increasing sample size.

<sup>8</sup>See Bernardo (1979a) for a thorough treatment and introduction of reference priors, and their relationship to other objective Bayesian approaches, such as Jeffreys (1939).

namely to affect a series of 0-1 outcomes generated by a randomly operating machine ( $\theta_0 = 0.5$ ) by means of alleged mental forces that would make the sample mean differ significantly from 0.5.

The null hypothesis holds that the results are generated by a machine operating with a chance of  $\theta = \theta_0 = 1/2$  in a Binomial model  $B(\theta, N)$ .<sup>9</sup> The alternative is the unspecified hypothesis  $\theta \neq \theta_0$ . The most evident test of that hypothesis is to observe a very long series of zeros and ones, which would give us enough evidence as to judge whether or not the null is compatible with the data.

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

$$z(x_1, \dots, x_N) = \frac{\sum_i x_i - N\theta_0}{\sqrt{N\theta_0(1-\theta_0)}} \approx 3.61 \quad (8)$$

and reject the null hypothesis on the grounds of the very low  $p$ -value it induces ( $p \ll 0.01$ ).

Compare this to the result of a proper 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 as  $B_{01}(x) \approx 19$  (cf. equation (6)). Using the zero-one loss model, he ends up accepting this hypothesis for a uniform prior probability over  $H_1$ . Hence, the data apparently support the null, in agreement with Lindley's diagnosis.

However, the BRC quantifies the expected loss from using  $\theta_0$  as a proxy for the true value  $\theta$  as substantial. Using a  $\beta(1/2, 1/2)$  reference prior for  $\theta$ , 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 2011, 18).

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. In other words, the theoretical reasons for conjecturing that there is something un-Bayesian about the BRC approach, especially with respect to using reference priors, have manifested themselves in a concrete case. On an un-charitable reading, the attempted Popper-Bayes reconciliation crumbles away as the Bayesianism of Bernardo's approach seems to be purely *instrumental*: that is, it makes use of Bayesian notation and assigns a "probability" over  $\theta$ , but ends up with conventional, automated inference procedures that recover frequentist results.

---

<sup>9</sup>The word "chance" need not refer to physical propensities or the like; we only need that these probabilities are intersubjectively agreed on.

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  $\theta$ : 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  $\hat{\theta}(x_1, \dots, x_n) = \bar{x}$  versus the null, we obtain (using the Normal approximation)

$$\begin{aligned} \log L_{\hat{\theta}, \theta_0}(x_1, \dots, x_N) &= \log \frac{p(\bar{x}|\hat{\theta})}{p(\bar{x}|\theta_0)} = \log \frac{p(x_1 = \hat{\theta}|\hat{\theta})^N}{p(x_1 = \hat{\theta}|\theta_0)^N} \\ &= \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^N - \log \left( \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-\frac{N}{2\sigma^2}(\hat{\theta} - \theta_0)^2} \right) \\ &= \frac{N}{2\sigma^2}(\hat{\theta} - \theta_0)^2 \xrightarrow{N \rightarrow \infty} \infty. \end{aligned} \tag{9}$$

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:  $\log L_{\hat{\theta}, \theta_0}(x_1, \dots, x_N) \approx 6.53$ ), in striking disagreement with the Bayes factor analysis.

Which Bayesian analysis is right, and which is wrong? In Jeffery's approach, two features are contentious, already touched upon in section 3. 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, we are not interested in 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.<sup>10</sup>

On the other hand, it may be argued that the reference Bayesian neglects, to the extent that she uses objective priors, crucial parts of our background knowledge. For example, the threshold for inferring to the presence of ESP should be higher than the threshold for inferring to the efficacy of a new pharmaceutical drug: if ESP were really efficacious, we would probably have found evidence for them before. A reference Bayesian thus needs to provide an account of how *statistical* (in)compatibility that needs to be calibrated to a *scientifically* meaningful judgment. This is, as I perceive it, still an open question. We can identify two possible levers: First, there is the utility constant  $d^*$  which we may interpret as the value of simplicity, informativity, explanatory power of the more specific null hypothesis in a specific context. Second, there are the (reference) prior distributions. As Bernardo suggests in his reply to Lindley (1999), we can calibrate reference priors with relevant prior information, e.g., assumptions about the variance of the prior over  $\theta$ . These “restricted” reference priors may be a sensible compromise between the need to incorporate context-sensitive information and the need for default distributions that can be used in scientific communication, and they illustrate the flexibility of BRC and the reference Bayesian framework.

A final concern is based on the observation that the likelihood ratio against the null, and the BRC score, reduce, for large data sets, to  $1/p(x|\theta_0)$  (cf. equation (9)). This quantity can be canonically related to the p-value under the null. Furthermore, Bayesian reference priors violate the Likelihood Principle (Berger and Wolpert 1984), one of the core principles of Bayesian inference. According to that principle, an inference must not depend on the statistical model of an experiment:

all the information about  $\theta$  obtainable from an experiment is contained in the likelihood function  $L_x(\theta) = P(x|\theta)$  for  $\theta$  given  $x$ . Two likelihood functions for  $\theta$  (from the same or different experiments) contain the same information about  $\theta$  if they are proportional to one another (Berger and Wolpert 1984, 19)

This principle is violated by Bernardo’s approach in the determination of reference priors. For instance, the Binomial and the Negative Binomial model induce the same likelihood function, but different reference priors. Thus, the reference priors depend on the stopping rule, which is an awkward phenomenon from a Bayesian point of view.

---

<sup>10</sup>This approach is, by the way, common practice in Bayesian model selection: When we are interested in predictive success of fitted models, we will often refrain from comparing two full-blown probability models, and compare their best representatives instead (e.g., Akaike’s AIC (Akaike 1973) or the Deviance Information Criterion DIC (Spiegelhalter et al. 2002)).

To my mind, this point need not worry us too much. The larger the sample size, the more information the data give us. If we are truly uncertain about the appropriate prior distribution, then we should adjust our demands for compatibility between the tested hypothesis and the data to the amount of available evidence. That is, the more informative the data, the more information can we put into the priors. Seen from the practice of hypothesis testing, the dependence of the priors on the probability model and the sample size does make sense.

Summing up, Bernardo's developments show that Bayesianism need not be identified with inferring highly probable models. That misconception, however, is perpetuated in post-Carnapian primers on Bayesian inference and has attracted Popper's understandable criticism. To give some more recent evidence: Howson and Urbach (1993, xvii) describe scientific reasoning as probabilistic reasoning, 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". We are now in a position to see that such purely probabilistic Bayesianism can at best be an interesting epistemic logic, but no appropriate model of scientific reasoning. Even in problems that are *prima facie* "purely epistemic", 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 philosophy of science, that is, confirmation theory.

Thus, if one retracts from classical subjective Bayesianism to a more general decision-oriented Bayesian model, the conflicts with Popper's falsificationist philosophy of inference diminish. The last section of the article wraps up our results and explores their implications for the Popper-Bayes relationship.

## 6 Conclusion: Popper and Bayes Reconciled

This paper has started out with two projects: First, to examine evidential interpretations of hypothesis tests. Here, hypothesis tests are not conceived of as practical decision tools for either accepting or rejecting a null hypothesis; rather, they serve as a basis for evidential judgments whether or not the null hypothesis is compatible with the data. Consequently, we have focused on the two-sided testing problem, involving a point null hypothesis  $H_0 : \theta = \theta_0$  embedded into some general alternative  $H_1 : \theta \neq \theta_0$ . The second project consisted in trying to relieve the tension between Popperian and Bayesian views on scientific method, testing, and inference. I believe that a solution of the first problem gives us a cue for the second one, too.

Since hypothesis testing emerged in frequentist statistics, it is fair to start

our investigation there. However, Fisher’s methodology of significance testing takes p-values/significance levels as a measure of disbelief of a hypothesis. That approach suffers from severe conceptual and technical shortcomings, and it is safe to say that p-values do not provide a basis for valid evidential assessments and inductive inferences.

These findings do not imply, however, that Bayesianism is the solution. The subjective, exclusively probabilistic Bayesian’s approach to hypothesis testing leads to counterintuitive results in the face of Lindley’s paradox, and also to foundational problems. The probabilist has a hard time recognizing any value in hypothesis tests, and fails to explain why they are an integral part of the practice of science.

The most satisfactory solution of this dilemma is, to my mind, provided by José Bernardo’s reference Bayesianism. There, the question of compatibility between data and null is subtly rephrased as the question of whether we can use  $\theta = \theta_0$  as a proxy for the more general model  $\theta \in \Theta$ . By adopting a proper decision-theoretic, forward-looking perspective that focuses on expected predictive success, on whether the observed effect will be replicated, the shortcomings of both the frequentist and the subjective Bayesian accounts are avoided, and a feasible explication of severely testing a precise null hypothesis is given. Although this methodology is usually combined with (objective) reference priors, the decision-theoretic developments by themselves do not require that one abandon a subjective take on prior probabilities, and become an objective Bayesian.

The title question should therefore be answered as follows: statistical hypothesis tests provide scientific evidence by stating whether we should consider a point null hypothesis as a representative of a more general model. The evaluation proceeds by means of balancing the expected predictive success, simplicity and informativity of either option – in line with Popper’s dictum that science aims at well-supported, but simple and testable models. From the reference Bayesian perspective, we can recognize the methodological value of testing hypotheses and explain why scientists shun, like Popper, inference to “highly probable” hypotheses.

Of course, there is a residual tension due to Popper’s misappreciation of the subjective probability interpretation. I should therefore add that this Popper-Bayes reconciliation refers, above all, to Popper as a *methodologist*, not as a metaphysician of probability. It should also be said that Popper was, as his corroboration measure (1) shows, open to approaches to measure empirical support in probabilistic terms.

Thus, Bernardo’s approach allows Popper and Bayes to live, if not as friends, in mutual respect: parts of Popper’s critical rationalism can be embedded into a Bayesian framework, or vice versa, some Bayesian techniques can assist the

critical rationalist. In the BRC account of hypothesis testing, the Bayesian’s “automatic” preference for more probable hypotheses is removed, and replaced by a more general decision-theoretic account. Therefore, testing along the lines of BRC might give the best statistical explication of empirical corroboration in a Popperian spirit, and be the best attempt to reconcile subjective Bayesian and frequentist thinking. It may well be a tenuous reconciliation, but these are often the most subtle, challenging and rewarding.

## References

- Akaike, Hirotugu (1973): “Information Theory as an Extension of the Maximum Likelihood Principle”, in: B. N. Petrov, F. Csaki (ed.), *Second International Symposium on Information Theory*, 267–281. Akademiai Kiado, Budapest.
- Albert, Max (2002): “Resolving Neyman’s Paradox”, *British Journal for the Philosophy of Science* 52, 69–76.
- Berger, James O. (2003): “Could Fisher, Jeffreys and Neyman Have Agreed on Testing?”, *Statistical Science* 18, 1–32.
- Berger, James O., and Mohan Delampady (1987): “Testing Precise Hypotheses”, *Statistical Science* 2, 317–352 (with discussion).
- Berger, James O., and Thomas Sellke (1987): “Testing a point null hypothesis: The irreconcilability of P-values and evidence”, *Journal of the American Statistical Association* 82, 112–139 (with discussion).
- Berger, James O., and Robert L. Wolpert (1984): *The Likelihood Principle*. Hayward/CA: Institute of Mathematical Statistics.
- Bernardo, José M. (1979a): “Reference posterior distributions for Bayesian inference”, *Journal of the Royal Statistical Society B* 41, 113–147 (with discussion).
- Bernardo, José M. (1979b): “Expected Information as Expected Utility”, *Annals of Statistics* 7, 686–690.
- Bernardo, José 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 (with discussion), Oxford University Press, Oxford.
- Bernardo, José M. (2011): “Integrated Objective Bayesian Estimation and Hypothesis Testing”, to appear in J. Bernardo et al. (eds.): *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*. Oxford University Press, Oxford.

- Bernardo, José M., and Adrian F. M. Smith (1994): *Bayesian Theory*. Chichester: Wiley.
- Carnap, Rudolf (1950): *Logical Foundations of Probability*. The University of Chicago Press, Chicago.
- Cohen, Jacob (1994): “The Earth is Round ( $p < .05$ )”, *American Psychologist* 49, 997-1001.
- Crupi, Vincenzo, Katya Tentori, and Michel Gonzalez (2007): “On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues”, *Philosophy of Science* 74, 229–252.
- Earman, John (1992): *Bayes or Bust?*. Cambridge/MA: The MIT Press.
- Fisher, Ronald A. (1935): *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Fisher, Ronald A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.
- Gillies, Donald (1971): “A Falsifying Rule for Probability Statements”, *British Journal for the Philosophy of Science* 22, 231–261.
- Good, I.J. (1952): “Rational Decisions”, *Journal of the Royal Statistical Society B* 14, 107–114.
- Harlow, L.L., S.A. Mulaik, and J.H. Steiger (eds.) (1997): *What if there were no significance tests?*. Mahwah/NJ: Erlbaum.
- Howson, Colin, and Peter 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 Scientific Exploration* 1, 21–50.
- Jefferys, W. H. (1990): “Bayesian Analysis of Random Event Generator Data”, *Journal of Scientific Exploration* 4, 153–169.
- Jeffreys, Harold (1939): *Theory of Probability*. Oxford: Clarendon Press.
- Joyce, James M. (1998): “A Nonpragmatic Vindication of Probabilism”, *Philosophy of Science* 65, 575–603.
- Lindley, Dennis V. (1957): “A statistical paradox”, *Biometrika* 44, 187–192.

- Lindley, Dennis V. (1999): “Discussion: Nested Hypothesis Testing: The Bayesian Reference Criterion”, in: J. Bernardo et al. (ed.), *Proceedings of the Sixth Valencia Meeting*, 122–124. Oxford University Press, Oxford.
- Mayo, Deborah G. (1996): *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, Chicago & London.
- Mayo, Deborah G., and Aris Spanos (2004): “Methodology in Practice: Statistical Misspecification Testing”, *Philosophy of Science* 71, 1007–1025.
- Mayo, Deborah G., and Aris Spanos (2006): “Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction”, *The British Journal for the Philosophy of Science* 57, 323–357.
- Neyman, Jerzy, and Egon Pearson (1933): “On the problem of the most efficient tests of statistical hypotheses”, *Philosophical Transactions of the Royal Society A* 231, 289–337.
- 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. (1956/83): *Realism and the Aim of Science*. Second Edition 1983. London: Hutchinson.
- Popper, Karl R. (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Harper.
- Popper, Karl R. (1972/79): *Objective Knowledge*. Oxford: Oxford University Press. Revised Edition.
- Royall, Richard (1997): *Scientific Evidence: A Likelihood Paradigm*. London: Chapman & Hall.
- Rowbottom, Darrell (2010): *Popper’s Critical Rationalism: A Philosophical Investigation*. London: Routledge.
- Rowbottom, Darrell (2011): “Popper’s Measure of Corroboration and  $P(h—b)$ ”, forthcoming in *British Journal for the Philosophy of Science*.
- Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde (2002): “Bayesian measures of model complexity and fit”, *Journal of the Royal Statistical Society B* 64, 583–639 (with discussion).
- Spielman, Stephen (1974): “On the Infirmities of Gillies’s Rule”, *British Journal for the Philosophy of Science* 25, 261–265.

Talbott, William (2008): “Bayesian Epistemology”, *Stanford Encyclopedia of Philosophy*, accessed on March 4, 2010 at <http://plato.stanford.edu/entries/epistemology-bayesian/>.

Ziliak, Stephen T., and Deirdre N. McCloskey (2004): “Size Matters: The Standard Error of Regressions in the *American Economic Review*”, *Journal of Socio-Economics* 33, 527–546.