RMM Vol. 3, 2012, 1–13
Special Topic: Statistical Science and Philosophy of Science
Edited by Deborah G. Mayo, Aris Spanos and Kent W. Staley
http://www.rmm-journal.de/

Jan Sprenger

The Renegade Subjectivist: José Bernardo's Reference Bayesianism*

Abstract:

This article motivates and discusses José Bernardo's attempt to reconcile the subjective Bayesian framework with a need for objective scientific inference, leading to a special kind of objective Bayesianism, namely *reference Bayesianism*. We elucidate principal ideas and foundational implications of Bernardo's approach, with particular attention to the classical problem of testing a precise null hypothesis against an unspecified alternative.

1. Introduction: A Stalemate Situation

The label 'objective Bayesianism' sounds almost self-contradictory. Bayesianism is a paradigm of inductive inference based on subjective degrees of beliefs, and seems to be anything but an objective form of inference. Therefore, objective Bayesianism cannot mean to deny *any* subjective element in inductive inference: rather, it aims at reconciling the subjective nature of Bayesianism with the aim of scientific objectivity.

Among the many approaches that are devoted to this end (for discussion, see Berger 2003; Mayo 2011), we focus on the *reference Bayesian approach*, mainly developed by José Bernardo over the last 30 years. The goal of the article consists in identifying, in a non-technical manner, the main elements and philosophical motivations of Bernardo's approach, as well as potential points of disagreement with subjective Bayesians and frequentists. We illustrate the implications of the reference Bayesian approach in the classical problem of testing a precise null hypothesis against an unspecified alternative.

In Bayesian inference, an agent's degrees of belief in hypotheses and theories conform to the axioms of probability. Beliefs are changed by Conditionalization, according to Bayes's theorem. In other words, Bayesianism is a theory of the *revision* of degrees of belief: it represents learning from experience and making inductive inferences as a rational belief updating process. The main requirements on that process are (1) that the function representing degrees of belief

* The author expresses his gratitude to the NWO for funding his research through the Veni grant "An Objective Guide for Public Policy? Scope and Limits of Data-Driven Methods in Statistical Model Evaluation" (016.104.079).

satisfy the axiom of probability functions and (2) that incoming evidence change the agent's degrees of belief in a hypothesis H according to Conditionalization on the evidence E:

$$P_{\text{new}}(H) := P(H|E) = P(H)\frac{P(E|H)}{P(E)}. \tag{1}$$

Thus, the end product of Bayesian inference is a posterior distribution, which represents our actual uncertainty after learning the evidence. The inferences that we make, and the decisions that we take, qualify as rational because they emerge as the result of a rational belief updating process.

However, it has been argued that scientific inference is very different from idealized scenarios for reasoning under uncertainty, such as drawing balls from an urn, or a game of chance, where Bayesian inference is obviously powerful. Practical problems such as computational costs put aside, scepticism vis-à-vis Bayesian methods in statistical inference can usually be traced to the following roots:

1. Scientific hypotheses are either false or true. Presumably, it is the task of science to state the *evidence* for a certain hypothesis, not to probabilify those hypotheses and to report degrees of belief in their truth. This way of thinking is typical of the *frequentist* approach in statistics, subsuming those ways of inference where probabilities are not interpreted as degrees of belief, but as relative frequencies of the occurrence of an event.[1] The intuition behind this objection is that science and statistics should investigate *objective* relations between phenomena and theories, between hypothesis and evidence.

2. Following up on the previous point, scientists often ask whether a certain hypothesis (e.g., X is independent of Y) is compatible with a given set of data, or whether a certain effect between quantities of interest is present. This is the basic question of statistical hypothesis tests. A Bayesian expresses this question by assuming a prior probability distribution and computing a posterior, but that seems to answer a different question. Moreover, it seems hard to convince a fellow scientist, a funding agency or the Food and Drug Administration that strong evidence can be counterbalanced by equally strong a priori 'prejudices'.

3. According to standard Bayesian theory, any system of degrees of belief that does not violate the axioms of probability and known empirical constraints counts as rational. In practice, this leaves ample space for the assignment of prior probabilities, and equally ample space for the resulting posteriors on which we base our decisions. If the result of a Bayesian experiment consists in a particular distribution of personal degrees of belief, what normative force do these results carry?

[1] This includes the works of Fisher, Neyman and Pearson as well as the more recent error-statistical approach of Mayo and Spanos 2006.

The justifications for being a subjectivist usually pertain to the decision-theoretic coherence of Bayesianism. Subjective Utility Theory, the dominant paradigm in decision theory, bases one's decisions on rational degrees of belief in various states of the world. By construction, the Bayesian approach nicely hooks up with decision theory. From the perspective of Expected Utility Theory, it can then be argued that several frequentist procedures, such the impact of stopping rules on inference, are decision-theoretically incoherent (e.g., Edwards et al. 1963; Kadane et al. 1996; Sprenger 2009).

Moreover, frequentist measures of evidence, such as p-values/significance levels, have been found to be poor measures of evidence in a variety of respects (Berger and Sellke 1987): evidence often serves as a justification to disbelieve or to give up a point hypothesis, but strong frequentist evidence such as a low p-value is actually a very poor guide to disbelief and rejection. Typically, the frequentist will overstate the evidence against the tested hypothesis compared to a subjective Bayesian analysis.

On the other hand, these advantages of Bayesianism carry less weight in circumstances where we cannot directly determine the material value of a right or wrong decision, e.g., in theoretical branches of science (Fisher 1956). Bayesian inference is basically a theory of what we should *believe*, and does not directly address certain inferential questions, such as 'what is the best estimate?', or 'how strong is the evidence?'. But these are the questions that many scientists are interested in.

So there seems to be a stalemate between the Bayesians, who are supported by decision theory, and the frequentists, who are supported by accepted scientific practice. Bernardo, who feels the pull of the arguments of either side, opts for de-subjectivizing the Bayesian account while at the same time maintaining its decision-theoretic foundation, which is arguably the greatest asset of Bayesianism. Moreover, he aims at a unification of estimation and hypothesis testing through a Bayesian lens. The next three sections expose the conceptual foundations of Bernardo's approach.

2. Intrinsic Loss Functions

One of the main problems of statistical inference is to estimate a quantity of interest on the basis of data x. A Bayesian will (if she is an expected utility maximizer) typically choose the *Bayes estimator* $\tilde{\theta}(x)$ for the loss function $L(\cdot,\cdot)$—the estimator that minimizes the expected loss on the basis of the posterior distribution:

$$\tilde{\theta}(x) = \text{argmin}_{\hat{\theta}\in\Theta} \int_{\Theta} L(\hat{\theta},\theta)\, p(\theta|x)\, d\theta \tag{2}$$

For instance, if the loss function L is the familiar quadratic loss $L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$, then the Bayes estimator amounts to the mean of the posterior distribution, that is: $\tilde{\theta}(x) = \int_{\Theta} \theta\, p(\theta|x)\, d\theta$.

However, Bayes estimators are not invariant under one to one transformations of the parameter space. That is, if we want to estimate $\psi = g(\theta)$, it will *not* hold in general that

$$\widetilde{g(\theta)}(x) = g(\tilde{\theta}(x)). \tag{3}$$

This has some counterintuitive consequences: when we want, as good Bayesians, to estimate the standard deviation of a random variable, then we cannot use a Bayes estimate of the variance and take the square root. This number will typically differ from a direct Bayes estimate of standard deviation, and vice versa. In other words, using a Bayes estimate of a certain quantity (such as variance) to infer to best estimates of other, canonically related quantities (such as standard deviation) is, from a Bayesian perspective, incoherent. This may be logical when we are dealing with monetary losses, but it is, according to Bernardo, "rather difficult to explain when, as it is the case in theoretical inference, we merely wish to report an estimate of some quantity of interest" (Bernardo 2011, 3), or in other words, when we are interested in the true value of θ.

The famous statistician R. A. Fisher made a similar point:

> "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)

In other words, the occurrence of loss functions in the Bayes estimator undermines the answer to a pure inferential problem, such as estimating the true value of θ. Decision-theoretically motivated approaches, such as going for the Bayes estimator, may be adequate for purposes of industrial quality control and the like, but not always for quantifying scientific evidence.

Fisher's criticism primarily aims at Neyman and Pearson's behavioral frequentist approach, but it equally applies to Bayesians since Bayes estimators depend, in general, on the assumed loss function. Bernardo is, on that point, in sync with Fisher, but the conclusions are different: unlike Fisher, he believes that Bayesian inference can be rescued: for good estimation in theoretical science, we have to work with loss functions that do not vary under one-to-one transformations.

Take the simple case of estimating the value of a parameter θ and measuring the loss that we suffer by working with θ_0 instead of the true parameter value. Bernardo recommends to switch from the distance between *parameters* to the distance between *models*:

> "It may naïvely appear that was is required is just some measure of discrepancy between θ_0 and θ. However, since all parametrizations are arbitrary, what is really required is some measure of the discrepancy between the *models* labeled by θ_0 and by θ. By construction, such a discrepancy measure will be invariant of the particular parametrization used." (Bernardo 2011, 6, original emphasis)

Following a suggestion by Robert (1996), Bernardo names these loss functions *intrinsic* since they measure the discrepancy between two probability models instead of the discrepancy between parameter values. The former, but not the latter, are invariant under one-to-one reparametrization. A natural choice for such a discrepancy measure between two distributions P_θ and P_{θ_0} with densities $p(\cdot|\theta)$ and $p(\cdot|\theta_0)$ is their mutual *Kullback-Leibler divergence* (Kullback and Leibler 1951) or *relative entropy*:

$$\delta(P_\theta, P_{\theta_0}) = \min\left(\int_{\mathscr{X}} p(x|\theta)\log\frac{p(x|\theta)}{p(x|\theta_0)}\,dx;\ \int_{\mathscr{X}} p(x|\theta_0)\log\frac{p(x|\theta_0)}{p(x|\theta)}\,dx\right). \quad (4)$$

This discrepancy is invariant under bijective transformations of the parameter space since the parameters affect the discrepancy only via the probability densities that they induce, which are independent of the particular parametrization.[2] Of course, there are also other divergence measures between probability distributions, but the logarithmic divergence is distinguished by a variety of theoretical virtues, and it has a straightforward anchoring in coding theory (Good 1952; Bernardo 1979b).

This particular loss function has, beside invariance, several other interesting properties which have been explored by Bernardo in a series of papers over the last 30 years. For example, it can also be calculated using a sufficient statistics $T(x)$ instead of the full data x. We skip the application-related developments, and only mention that the loss function in (4) can be interpreted as the expected minimal log-likelihood ratio in favor of the true model. Thus, the intrinsic loss function does not only have the desired invariance property: also from a frequentist point of view, it is related to relevant measures of evidence (e.g., in the Neyman-Pearson Lemma).

3. Reference Priors

One of the biggest problem for the Bayesian consists in developing a sound and practicable methodology for assigning prior distributions: "in many situations however, either the available prior information on the quantity of interest is too vague to warrant the effort required to formalize it, or it is too subjective to be useful in scientific communication" (Bernardo 2011, 10).

The second pillar of Bernardo's reference Bayesianism, next to the development of invariant loss functions, consists in an appropriate choice of reference prior distributions. While in classical, subjective Bayesianism, the prior distribution reflects one's prior degrees of belief, a reference Bayesian recognizes the difficulty of coming up with a meaningful subjective priors in a variety of problems. Therefore she resorts to a conventional default choice.

Bernardo's key idea for selecting such reference priors consists in *maximizing the information of the data*, that is, in maximizing the information that the

[2] Taking the minimum is necessary since KL-divergence is not symmetric.

data transmit about the parameter of interest. This idea is most easily illustrated in a one parameter model $\mathscr{M} = \{p(x|\theta), \theta \in \Theta, x \in \mathscr{X}\}$, with parameter space Θ and sample space $\mathscr{X}$. The information in the data is explicated as the expected Kullback-Leibler discrepancy between the prior probability density $p(\theta)$ and the posterior probability density $p(\theta|x)$ (Bernardo 1979a, 114–115):

$$I^{\Theta}(\mathscr{M}, p(\theta)) = \mathbb{E}_x \left[\int_{\Theta} p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)} \, d\theta \right] \tag{5}$$

Now consider this quantity as the experiment is replicated an indefinite number of times. Instead of a single realization x, we then deal with data $x_1, \ldots, x_k$. As $k \to \infty$, the functional $I^{\Theta}(x^k, p(\theta))$ will approach the amount of information about θ which is missing in the prior $p(\theta)$. The reference prior or 'noninformative' prior distribution is the probability distribution that maximizes $I^{\Theta}(x^k, p(\theta))$ as $k \to \infty$, that is, that makes the data maximally informative.

In this construction, the reference prior depends on the particular model that has been specified. For example, the reference prior over a finite sample space is the uniform distribution, and the reference prior over a parameter $\theta \in [0, 1]$ in a binomial model of the data will be $Be(1/2, 1/2)$-distributed ($p(\theta) \propto 1/\sqrt{\theta(1-\theta)}$), whereas for the closely related *Negative* Binomial model, whose only difference to the Binomial model concerns the sampling rule, the reference prior will look differently.

It follows that inference (e.g., estimation) with reference priors leads to different answers, depending on which probability model we use. In particular, inference depends on the sampling rule that has been used for obtaining a data set, violating the *Likelihood Principle* (Berger and Wolpert 1984), one of the core principles of subjective Bayesian inference. According to that principle,

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

This principle is violated in Bernardo's objective Bayesianism since, as he frankly admits, the inference depends via the reference priors on the specified probability model. Indeed, violation of the Likelihood Principle leads to counterintuitive consequences, similar to those that frequentists experience. Assume, for instance, that we are repeating Bernoulli experiments until either a certain number of successes ($s = 10$), or a certain number of trials ($N = 100$), has been occurred. What should be our reference prior if we observe the 10th success on the 100th repetition? The one for the Binomial distribution? The one for the negative Binomial distribution? A mixture of both (Geisser 1984)?

Such observations highlight (i) that with respect to foundations, reference Bayesians deviate substantially from orthodox subjective Bayesian inference, and (ii) that they need a defence against the arguments that have been made in favor of the Likelihood Principle. This is still an open topic of debate—see the

replies and rejoinder to Bernardo (2011). One answer could be that reference priors are motivated by the desire to make the data as informative as possible, compared to the information in the priors. 'Information' is, in the above definition, once again explicated by means of Kullback-Leibler divergence. Reference priors are then referential, or so one could argue, because for an unbiased judgment in the absence of meaningful prior information, it is certainly useful to start with as little information as possible, or to give as much weight to the data as possible, and to minimize the impact of prior opinion. That is the sense of default priors that reference priors explicate. One may disagree about the adequacy of Bernardo's specific proposal, but it remains a coherent way of defining 'noninformative priors' that does not appeal to philosophically dubious symmetry and indifference principles.

Moreover, reference priors need not be understood as the *solution* to an inference problem, but rather, as the name suggests, as a *reference point* against which we can gauge the results of a subjective Bayesian analysis (cf. Bernardo 1997). Reporting a reference analysis alongside a subjective analysis enables us to assess the strength of our subjective presumptions. So reference priors can also be understood as a form of *sensitivity analysis*. When we are working with subjective priors, we would like to know to what extent our conclusions are sensitive to our choice of the prior. Understanding reference priors as a gauging instrument for our subjective beliefs, as a possible common ground for debating between statisticians with different opinions strips them off the ambition to figure as uniquely rational degrees of belief.

Keeping this in mind, we can now proceed to the objective Bayesian's approach to the problem of testing a point null hypothesis. After going through that treatment, we will be in a better position to assess the overall merits and drawbacks of the reference Bayesian approach.

4. Hypothesis Testing

One of the core ambitions of Bernardo's reference Bayesianism is a unified approach to hypothesis testing and estimation: where standard, subjective Bayesian approaches to hypothesis testing make an accept/reject decision on the basis of the consequences of a wrong decision, the reference Bayesian aims at answering the question whether or not the parameter value θ_0 is compatible with the data (Bernardo 1999). These questions are different, and it is not clear whether it even makes sense to ask the latter in a subjective Bayesian framework—after all, there is no evidence independent of subjective belief. In the rest of this section, we will elaborate the differences between classical frequentist, subjective Bayesian and reference Bayesian methods regarding testing a point null hypothesis for compatibility with the data.

We consider the standard problem of testing the mean of a normal distribution with known variance σ^2, where the null has the form $H_0 : X \sim \mathcal{N}(\theta_0, \sigma^2)$, and the alternative has the unspecified form $H_1 : X \sim \mathcal{N}(\theta, \sigma^2)$ with $\theta \neq \theta_0$. For

large samples, a frequentist tester will reject the null hypothesis if the absolute value of conventional deviance statistic $z(X_1,\ldots,X_N) := (\sum_k X_k - N\theta_0)/(\sqrt{N}\sigma)$ exceeds a certain threshold. This practice is justified by the observation that z converges, due to the Central Limit Theorem, in distribution to the standard Normal distribution $\mathcal{N}(0,1)$. A frequentist significance tester rejects the null if the results are 'very unexpected' under H_0, that is, if they are in the far tails of the distribution of z; otherwise, the null is judged to be compatible with the data.[3]

A subjective Bayesian takes a different stance. This can be illustrated in the asymptotic case $N \to \infty$. If we fix, for increasing sample size N, the significance level of the data—and let the data be highly significant against the null—, then the Bayes Factor in favor of the null (that is, the ratio of posterior and prior odds) will exceed any bound, and the posterior probability of the null hypothesis will converge to 1. This phenomenon, known as the the Jeffreys-Lindley paradox (Lindley 1957) is remarkable because it shows a fundamental difference between the type of results that a Bayesian and a frequentist obtain from their respective data analysis.[4] Formally:

> **Lindley's Paradox:** Take a Normal model $N(\theta,\sigma^2)$ with known variance σ^2, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, assume $p(H_0) > 0$ and any regular proper prior distribution on $\{\theta \neq \theta_0\}$. Then, for any testing level $\alpha \in [0,1]$, we can find a sample size $N(\alpha, p(\cdot))$ and independent, identically distributed (i.i.d.) 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)

Thus, Bayesians contend that the value of $z(x)$ is no reliable indication of the tenability of the null hypothesis. This is not surprising: If we are testing the null hypothesis 'for real', then we also assign a proper prior probability to H_0. That is, we give H_0 a non-zero probability of being true, e.g., $P(H_0) = \varepsilon > 0$—the precise value does not matter—with the remaining probability mass spread out over the real line. In this case, it makes sense that 'significant' results nevertheless favor H_0 over H_1: For a 'significant difference' between θ_0 and the averaged, variance-corrected sample mean z becomes the smaller, the more N increases. In particular, this deviance will appear small compared to the deviance to most of the hypotheses that are part of H_1. In other words: as soon as we take our priors seriously, as a honest expression of our subjective uncertainty, it is clear that we will end up with results favoring θ_0 over an unspecified alternative. On that reading, the 'paradox' just demonstrates that statistical significance is a

[3] Of course, there are also more sophisticated approaches such as Mayo and Spanos' (2006) error-statistical approach, but a comparison of reference Bayesianism to their framework would go beyond the scope of this article.

[4] True, the frequentist confidence interval around θ_0 will become more and more narrow, but for the moment, we would like to *test* the null hypothesis, not to find an interval estimate.

poor indicator of rational belief. Indeed, it sounds absurd that minute deviations of the sample mean from θ_0 provide significant evidence against the null if the sample is just large enough!

However, the Bayesian has made a substantial presumption: namely that we can assign a meaningful singular prior $\varepsilon > 0$ to the null hypothesis. But in practice, this need not be the case. Consider testing the efficacy of a medical drug, the palate of a wine taster, or the bias of a measuring instrument: there is nothing that distinguishes θ_0 (in terms of degrees of belief) from other values of θ that are in its immediate neighborhood. Therefore a singular prior seems inadequate. It is questionable whether such a subjective Bayesian analysis really answers the questions 'can I use the parameter value θ_0 as a proxy for the unknown true value of θ?' or 'are the data incompatible with $\theta = \theta_0$'? Subjective Bayesians presuppose that there is something special about the value of θ_0, but this will typically not be the case.[5]

Bernardo's solution to this dilemma is arguably ingenious: he understands a hypothesis testing problem as a proper decision problem where we decide on whether or not to treat the data as generated by the null hypothesis $\theta = \theta_0$. In other words, the inferential problem becomes a *decision problem* where the different options (accept/reject H_0) are judged according to their expected utilities. The utility of accepting H_0 when it is wrong is, however, not given externally, but quantified by means of the expected predictive score of using θ_0 instead of the true value θ. This score is calculated by averaging the logarithmic score $(-\log p(x|\theta_0))$ over the sample space $\mathscr{X}$.

Cutting short the technicalities here, Bernardo (1999) manages to show that under these presumptions, the difference in expected utilities between accepting and rejecting the null is (leaving nuisance parameters aside) essentially a function of the term

$$\int \delta_0(P_{\theta_0}, P_\theta) = \int p(\theta|x) \left(\int p(y|\theta) \log \frac{p(y|\theta)}{p(y|\theta_0)} \, dy \right) d\theta. \tag{6}$$

In other words, the decision in a hypothesis testing problem depends, according to Bernardo, on the data via the expected intrinsic loss (6), cf. equation (4). Anticipating our final evaluation, we can see from (6) that Bernardo has accomplished an unified account of estimation and hypothesis testing: both treatments crucially depend on the intrinsic loss or the intrinsic distance between the estimated/hypothesized parameter value and the true parameter value.

Coming back to hypothesis testing proper, this Bayesian Reference Criterion (BRC, Bernardo 1999, 108) allows for some interesting results. Assume, for instance, that we are testing a subject that claims to possess extrasensory capaci-

[5] It does not help either to cite the results of Berger and Delampady 1987. Their Theorem 1 demonstrates that the testing of a point null hypothesis with non-zero prior can be understood as a convenient simplification of testing the null hypothesis $|\theta - \theta_0| \leq \varepsilon$ for a small value of ε, with a continuous, but rather sharply peaked prior. More precisely, the Bayes factors in both testing problems can be related to each other. Unfortunately, this approximation breaks down as N increases, making this justification unavailable to the subjective Bayesian hypothesis tester in the situation of Lindley's paradox.

ties, namely to affect 0-1-outcomes generated by a randomly operating machine by means of mysterious mental forces. Recording the number of zeros and ones, we check whether there is a significant difference between them. For a large sample, it turns out that there have been 52.263.471 zeros in 104.490.000 trials (Jahn et al. 1987). A proper Bayesian (Jefferys 1990) assigns a non-zero probability to the null hypothesis and accepts this hypothesis for most choices of prior probabilities over H_1 (the Bayes factor is, for a typical choice, about 19 in favor of H_0). However, a reference Bayesian will conclude that the expected loss from using θ_0 as a proxy for the true value θ (= the expected log-likelihood ratio against H_0) is substantial, namely $\log 1400 \approx 7.24$. This is in sync with the observation that any non-dogmatic prior yields the posterior $\theta \sim N(0.50018, 0.000049)$, where the low variance establishes that "under the accepted conditions, the precise value $\theta_0 = 1/2$ is rather incompatible with the data" (Bernardo 2011, 18). This does, of course, not prove the extrasensory capacities of our subject; a much more plausible explanation is a small bias in the random generator.

The crucial difference to a subjective, Bayes factor analysis is that the entire space of alternatives is *not* integrated out. Recall that that analysis favored the null over the alternative because it 'defeated' all parameter values that were far from θ_0, irrespective of whether θ_0 or a different value close to it was the actual truth. Whereas the objective Bayesian restricts herself to pointing out that the evidence against θ_0 with respect to *some* other value is very strong. In that sense, while avoiding the integrating-out or catch-all treatment that is characteristic of subjective Bayesianism, the reference Bayesian approach is built on very Bayesian grounds: namely maximizing Subjective Expected Utility. The last section proposes some subjective Bayesian rejoinders to that view, and summarize our findings.

5. Conclusion: A Critical Appraisal

This short paper has presented the three main elements of Bernardo's reference Bayesian approach: intrinsic loss functions for estimation, reference priors, and a decision-theoretic, prediction-oriented access to hypothesis testing. Strictly speaking, these elements are independent, but they fit into a coherent philosophy of inference because they are related in a number of ways. For instance, the (expected) logarithmic score $-\log p(\cdot)$ is used in all three elements: it quantifies missing information about θ, expected divergence between prior and posterior, and predictive success of a hypothesis. Similarly, the intrinsic loss function for estimation problems also figures in the Bayesian Reference Criterion (BRC) for hypothesis testing.

How should we place this framework on a scale between Bayesian and frequentist approaches? When stressing the difference between objective and subjective Bayesianism, it should be noted that the reference prior approach does not only provide a unified approach to testing and estimation problems: it also lays a decision-theoretic foundation for testing point null hypotheses. Although

the test statistics are the same like in frequentist inference, making the method attractive for a non-Bayesian as well, Bernardo's method is distinguished by its decision-theoretic foundation. This is a big asset vis-à-vis frequentist philosophies of induction that are based on p-values or mathematical derivatives thereof (Berger and Delampady 1987; Berger and Sellke 1987). Belief is now separated from evidence. However, three fundamental challenges deserve mention. It would go beyond the scope of the paper to try to answer them here; we direct the interested reader to Bernardo (1999; 2011) and Lindley (1972).

1. Does it really make sense to consider hypothesis testing problems as stylized prediction problems? Is this really adequate and sufficiently general for modeling inferences in theoretical science? Aren't we reducing the complexity of real hypothesis tests to a scheme that does not necessarily fit them?

2. Lindley (1999) objects, in his discussion of Bernardo's (1999) paper, that a crucial advantage of Bayesianism gets lost in the Bernardian synthesis of frequentist and Bayesian techniques: *context-sensitivity*. Compare the testing of the efficacy of a new medical drug to the testing the extrasensory capacities of an arbitrary subject. Clearly, we will be much more willing to reject the null hypothesis (no efficacy) in the first case than in the second case: quite often, new drugs turn out to be effective whereas if extrasensory capacities really existed, we would probably have observed them in previous experiments. A proper Bayesian would therefore, in the second case, use a prior that is much more spiked around θ_0 than in the second case. The disadvantage of Bernardo's approach is, according to Lindley, that by the automatic use of the reference prior machinery, we deprive ourselves of the chance to distinguish between those cases. We are dealing with 'Greek letters' and dismiss our foreknowledge and scientific judgment. In his rejoinder, Bernardo (1999) proposes the use of appropriately restricted reference priors to accommodate Lindley's reservations.

3. The difference between statistical and scientific significance. Bernardo's conclusion that in the ESP example of the previous section, the null is incompatible with the data, might be true from a purely statistical point of view. But *scientifically*, it seems that the data vindicate the null because the observed effect is too small to be the product of interesting extrasensory capacities. Probably it is just an artefact of the sampling device. The question is then: what use does it have to say that the data are (statistically) incompatible with the null hypothesis when this does not help us to decide whether or not we should see the null as (scientifically) confirmed?

Summing up, the gist of Bernardo's reference Bayesianism is not to replace subjective inference in science and statistics. Rather, it is an extension of the Bayesian machinery to cases where a proper subjective analysis is not feasible for whatever reasons. Independent of one's stance towards this project, Bernardo deserves credit for coming up with an account of reference priors that

unifies estimation and testing problems on information-theoretic and decision-theoretic grounds, and that has been fruitfully applied to a variety of non-trivial statistical problems.

References

Berger, J. O. (2003), "Could Fisher, Jereys and Neyman Have Agreed on Testing?", *Statistical Science* 18, 1–32 (with discussion).

— and J. M. Bernardo (1992), "On the Development of Reference Priors", in: Bernardo, J. M., J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 4*, Oxford: Oxford University Press, 35–60 (with discussion).

— and M. Delampady (1987), "Testing Precise Hypotheses", *Statistical Science* 2, 317–352 (with discussion).

— and T. Sellke (1987), "Testing a Point Null Hypothesis: The Irreconciliability of P-values and Evidence", *Journal of the American Statistical Association* 82, 112–139 (with discussion).

— and R. L. Wolpert (1984), *The Likelihood Principle*, Hayward: Institute of Mathematical Statistics.

Bernardo, J. M. (1979a), "Reference Posterior Distributions for Bayesian Inference (with discussion)", *Journal of the Royal Statistical Society* B 41, 113–147.

— (1979b), "Expected Information as Expected Utility", *Annals of Statistics* 7, 686–690.

— (1997), "Noninformative Priors Do Not Exist: A Discussion", *Journal of Statistical Planning and Inference* 65, 159–189.

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

— (2011), "Integrated Objective Bayesian Estimation and Hypothesis Testing", in: Bernardo, J. M. et al. (eds.), *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*, Oxford: Oxford University Press, 1–68.

Edwards, W., H. Lindman and L. J. Savage (1963), "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70, 450–499.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

— (1956), *Statistical Methods and Scientific Inference*, New York: Hafner.

Geisser, S. (1984), "On Prior Distributions for Binary Trials", *American Statistician* 38, 244–251.

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

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

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

Kadane, J., M. Schervish and T. Seidenfeld (1996), "Reasoning to a Foregone Conclusion", *Journal of the American Statistical Association* 91, 1228–1235.

Kullback, S. and R. A. Leibler (1951), "On Information and Sufficiency", *Annals of Mathematical Statistics* 22, 79–86.

Lindley, D. V. (1957), "A Statistical Paradox", *Biometrika* 44, 187–192.

— (1972), *Bayesian Statistics: A Review*, Philadelphia: SIAM.

— (1999), "Discussion: Nested Hypothesis Testing: The Bayesian Reference Criterion", in: Bernardo, J. M. et al. (eds.), *Proceedings of the Sixth Valencia Meeting*, Oxford: Oxford University Press, 122–124.

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

— (2011), "Statistical Science and Philosophy: Where Do/Should They Meet in 2011 (and Beyond)?", forthcoming in *Rationality, Morality and Markets*.

— and A. 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.

Robert, C. P. (1996), "Intrinsic Loss Functions", *Theory and Decision* 40, 192–214.

Sprenger, J. (2009), "Evidence and Experimental Design in Sequential Trials", *Philosophy of Science* 76, 637–649.