Statistica Sinica **3**(1993), 601-608

A NOTE ON JEFFREYS-LINDLEY PARADOX

Christian P. Robert

Université de Rouen and Purdue University

Abstract: The Jeffreys-Lindley paradox, namely the fact that a point null hypothesis will always be accepted when the variance of a conjugate prior goes to infinity, has often been argued to imply prohibiting the use of improper priors in hypothesis testing. We reevaluate this paradox by considering the role of the prior hypothesis probabilities and obtain a noninformative answer which is equivalent decisionwise to the classical *p-value.*

Key words and phrases: Hypothesis testing, vague prior, noninformative answer, *p*-value.

1. Introduction

In hypothesis testing, it is well-known that Bayesian and frequentist answers may drastically differ. For instance, Berger and Sellke (1987) and Berger and Delampady (1987) have shown that the smallest posterior probability of a point null hypothesis is usually much larger than the corresponding frequentist answer, the *p-value.* Lindley (1957) shows that the disagreement may be dramatic, in the following sense. Let $x \sim \mathcal{N}(\theta, 1)$ and the null hypothesis to test is $H_0 : \theta = 0$. If one uses conjugate priors, $\theta \sim \mathcal{N}(0, \sigma^2)$, with prior probability ϱ_0 for the null hypothesis H_0, the posterior probability of H_0,

$$\left[1 + \frac{1-\varrho_0}{\varrho_0} \frac{e^{-x^2/2(\sigma^2+1)}}{e^{-x^2/2}} \frac{1}{\sqrt{\sigma^2+1}}\right]^{-1}, \tag{1.1}$$

goes to 1 as σ^2 goes to infinity, whatever ϱ_0 and x are.

This result is statistically paradoxical because, first, a large value of σ^2 somehow corresponds to a noninformative setup and, therefore, answers from noninformative priors seem to be useless in this problem. Secondly, it is usually the case in estimation settings that the limit of conjugate estimators is equivalent to a "classical" frequentist answer and this property does not seem to occur for hypothesis testing. Obviously, the fact that (1.1) goes to 1 is not a mathematical paradox since the prior sequence is giving less and less mass to any neighborhood of 0 as σ^2 goes to infinity.

Many authors have commented on this paradox, either to criticize the Bayesian approach (Shafer (1982)) or to dismiss the use of improper priors for testing (Jeffreys (1961) and DeGroot (1982)). According to Berger (1991), it shows that a noninformative answer is not possible in this context and moreover, that it is in accordance with Occam's razor rule, i.e. that between two equally likely explanations, we should always choose the simplest one if no additional argument supports the other one (Berger and Jeffreys (1991)). A recent and detailled discussion of the Jeffreys-Lindley paradox is provided by Aitkin (1991).

However, recent decision-theoretic considerations of the testing problem in Hwang et al. (1992) have shown that improper priors were definitely necessary. For instance, in the Jeffreys-Lindley setup, the p-value $p(x) = 2(1 - \Phi(|x|))$ is inadmissible under squared-error loss,

$$(\Pi_0(\theta) - p(x))^2,$$

where Π denotes the indicator function, but cannot be dominated by a proper Bayes estimator, i.e. a true posterior probability. Furthermore, generalized Bayes answers, i.e. solutions of the form

$$\frac{\varrho_0 \varphi(x)}{\varrho_0 \varphi(x) + (1-\varrho_0)\int_{-\infty}^{+\infty} \varphi(x-\theta)\pi_1(\theta)d\theta} \tag{1.2}$$

where φ is the standard normal density, π_1 is a σ-finite measure and ϱ_0 the prior probability of H_0, are also admissible under squared-error loss and form a *minimal complete class.* Moreover, the least favorable answers – lower bounds on the posterior probabilities of H_0 – obtained in Berger and Sellke (1987) correspond formally to noninformative procedures of the form (1.2), as shown by Robert and Caron (1991).

As pointed out by DeGroot (1982), the trouble with using improper priors is that if one replaces the σ-finite measure $\pi_1(\theta)$ by the rescaled measure $c\pi_1(\theta)$, the constant c can be chosen to give any desired answer. We will show in the next section that, nonetheless, there exists a way to obtain the "proper" constant c for the Jeffreys prior by considering again a sequence of conjugate priors. The resulting noninformative answer is then no longer uniformly equal to 1 and, furthermore, provides an estimator which is surprisingly close to the classical p-value (for most decision purposes). Aitkin (1991) and Smith and Spiegelhalter (1980) also propose alternative techniques to select the constant c or remove the choice of this constant. The next section presents these answers and compares them with our solutions.

2. Reweighting the Alternatives

The fundamental argument underlying our reevaluation of the Jeffreys-Lindley paradox is that the prior probability ϱ_0 of the null hypothesis H_0 should depend on the prior variance under the alternative hypothesis H_1, σ^2. Such a dependence may seem absurd at first but consider that, from a Bayesian point of view, we are actually testing $H_0 : \theta = 0$ versus $H_1 : \theta \sim \mathcal{N}(0, \sigma^2)$. Therefore, the prior probability of H_1 (and therefore of H_0) may vary with σ^2. Indeed, while taking $\varrho_0 = 1/2$ is seemingly the fairest (or the most objective) choice, it does not take into account the fact that the alternative prior π_1 considers a larger set of possible values of θ as σ^2 increases, i.e. that the "effective support" of π_1 (say, the 99% HPD region) is getting larger as σ^2 goes to infinity. Larger values of σ^2 do not exclude smaller values of θ but, on the contrary, increase the range of values of θ compatible with H_1. In this sense, an increasing sequence of σ^2 leads to a sequence of *imbedded* alternative hypotheses. Therefore, the prior probability of H_1 should *increase* with σ^2. Such a dependency is also justified if we look at it the other way: a restriction of the range of possible values for θ under H_1 can result from some observations which are incompatible with the previous range of π_1 and which, therefore, partially argue against H_1. It is thus coherent to lower the prior probability of H_1 when the range of π_1 is decreasing. It is because ϱ_0 is kept constant that the Jeffreys-Lindley paradox occurs; we have to prevent the alternative prior mass from going to $\pm\infty$ too quickly. Casella and Berger (1987) noted that $\varrho_0 = 1/2$ was "too large" but did not pursue the reasoning leading to a prior dependent ϱ_0.

A natural requirement on the sequence of priors is that they should give sufficient weight to the range of values of θ which actually caused H_0 to be tested, i.e. to the θ's in a neighborhood of 0 which generate x's which could also originate from a $\mathcal{N}(0,1)$ distribution. Since, for σ large enough and a arbitrary, we have

$$\begin{aligned}\pi([-a,0) \cup (0,a]) &= (1-\varrho_0)[\Phi(a/\sigma) - \Phi(-a/\sigma)] \\ &\simeq (1-\varrho_0)\frac{2a}{\sigma}\varphi(0),\end{aligned}$$

it seems reasonable to impose the following restriction on ϱ_0, thus denoted by $\varrho_0(\sigma)$, to stress its dependency on σ,

$$\frac{1-\varrho_0(\sigma)}{\sigma} = c,$$

where c is a constant to be determined.

However, this constraint is too strong to hold when σ goes to infinity, since the prior probability of any fixed interval must go to 0. A more realistic requirement is therefore to choose $\varrho_0(\sigma)$ in such a way that the ratio of the prior probability of the null hypothesis to the prior probability of the "reasonable" range, $[-a,0)\cup(0,a]$, remains constant as σ goes to infinity, i.e.

$$(1-\varrho_0(\sigma))[\Phi(a/\sigma)-\Phi(-a/\sigma)] \propto \varrho_0(\sigma). \tag{2.1}$$

For σ large enough, this condition leads to the following equation

$$\frac{1-\varrho_0(\sigma)}{\sigma} \propto \varrho_0(\sigma). \tag{2.2}$$

In order to completely determine the dependency of ϱ_0 on σ^2, i.e. the proportionality factor in the above relation, we consider that 0 should have the same weight under both alternatives, namely that the densities are equal at 0,

$$\varrho_0(\sigma) = (1-\varrho_0(\sigma))\frac{1}{\sqrt{2\pi}\sigma}. \tag{2.3}$$

This implies that 0 is "indifferent" under both alternatives, whatever σ is. Note that, under this constraint, $\varrho_0(\sigma)$ goes to 0 when σ^2 goes to infinity. Such a behavior was also observed by Bernardo (1980) when implementing the reference prior approach in this setting. The posterior probability associated with (2.3) is then

$$\left[1+\sqrt{\frac{\sigma^2}{\sigma^2+1}}\sqrt{2\pi}\,e^{\sigma^2x^2/2(\sigma^2+1)}\right]^{-1}, \tag{2.4}$$

which converges to

$$(1+\sqrt{2\pi}\,e^{x^2/2})^{-1} \tag{2.5}$$

when σ^2 goes to infinity. Note that ϱ_0 in (2.3) converges to 1 when σ^2 goes to 0, as it should since H_0 is then true a priori, while $\varrho_0 = 1/2$ leads to a posterior probability of 1/2 in (1.1).

A most interesting feature of (2.5) is that it also corresponds to the generalized Bayes answer associated with the Jeffreys prior $\pi_1(\theta) = 1$ and $\varrho_0 = 1/2$, i.e.

$$\pi(\theta) = \frac{1}{2}\{\Pi_0(\theta)+\pi_1(\theta)\}$$

(where Π_0 denotes the Dirac mass at 0), since (1.2) leads to

$$\frac{e^{-x^2/2}/\sqrt{2\pi}}{e^{-x^2/2}/\sqrt{2\pi}+1}$$

in this case. Therefore, when the prior probability of H_0 depends on the prior variance σ^2, the Jeffreys estimator is the limit of the conjugate answers, as it is for point estimation. Moreover, this result indicates that $c = 1$ is the proper constant in this case. The equiponderance device (2.3) thus allows the determination of the effective constant for the Jeffreys prior.

Spiegelhalter and Smith (1982) and Aitkin (1991) obtained, respectively, a corresponding constant for the normalization of the improper prior π_1 by using *virtual* observations or by using *twice* the actual observations. In fact, Aitkin (1991) eliminates the problem of selecting a constant by replacing the Jeffreys prior π_1 by the posterior distribution $\pi_1(\theta|x)$ which is a $\mathcal{N}(x, 1)$ distribution in this case. The corresponding posterior probability of H_0 is

$$\left(1 + e^{x^2/2}/\sqrt{2}\right)^{-1}, \tag{2.6}$$

which leads to a constant $c = 1/\sqrt{2}$ for the Jeffreys prior. The main problem with this *ad hoc* solution is that it does not belong to the Bayesian paradigm because of the repeated use of the observations and that it lacks *coherency* as pointed out in the discussion following the paper (see, e.g., Lindley (1991)).

The method proposed in Spiegelhalter and Smith (1982) can be considered as determining the constant c for which the most favorable observation, i.e. $x = 0$, would give a Bayes factor of 1. In this case, it is $c = \sqrt{2\pi}$, which gives the posterior probability (for $\varrho_0 = 1/2$)

$$\left(1 + e^{x^2/2}\right)^{-1}; \tag{2.7}$$

this quantity also corresponds to the lower bound of Berger and Sellke (1987) and approximately to the answer associated with an uniform prior on $[-1.25, 1.25]$. It thus seems difficult to advocate the use of this precise proper prior as a noninformative prior. Similarly, Smith and Spiegelhalter (1980) point out that the elimination of the Jeffreys-Lindley paradox relies on a sufficient weighting of a neighborhood of $\theta = 0$ under H_1; however, they propose the specific proper priors, $\mathcal{N}(0, 1)$ and $\mathcal{U}_{[-1.96,1.96]}$ which lead, respectively, to the drastically different posterior probabilities

$$\left(1 + e^{x^2/4}/\sqrt{2}\right)^{-1} \quad \text{and} \quad \left(1 + e^{x^2/2}e^{-3/4}\right)^{-1}.$$

(For instance, for $x = 1.96$, the first probability is 0.35 and the second one 0.24, to be compared with the much lower values of Table 1.)

3. The Resulting Noninformative Answer

The dependency of ϱ_0 on σ^2 thus avoids the undesirable convergence to 1 and provides an estimator which can be considered as a noninformative answer and a Bayesian counterpart to the p-value. However, the validity of our derivation may be questioned, since the limiting prior resulting from (2.3) also has some undesirable features. Actually, for every $\varepsilon > 0$, one has

$$\pi([-\varepsilon, \varepsilon]) = \varrho_0(\sigma) + (1 - \varrho_0(\sigma))[\Phi(\varepsilon/\sigma) - \Phi(-\varepsilon/\sigma)],$$

where Φ is the standard normal cdf. Given (2.3), we get

$$\begin{aligned}\pi([-\varepsilon, \varepsilon]) &= \left[\frac{1}{\sqrt{2\pi}\sigma} + \Phi(\varepsilon/\sigma) - \Phi(-\varepsilon/\sigma)\right]\left(1 + \frac{1}{\sqrt{2\pi}\sigma}\right)^{-1} \\ &= \frac{1}{1+\sqrt{2\pi}\sigma}\Big(1 + \sqrt{2\pi}\sigma[\Phi(\varepsilon/\sigma) - \Phi(-\varepsilon/\sigma)]\Big),\end{aligned}$$

which converges to 0 as σ^2 goes to infinity. Therefore, the limiting prior gives no positive probability to any neighborhood of 0, and this behavior seems to be quite unreasonable. But this is usually the case with improper priors: they cannot be handled in the same way as subjective priors and, as pointed out by DeGroot (1982), *they should not be regarded as representing ignorance.* This feature of improper priors is present in most statistical problems and, therefore, should not prevent us from considering (2.5) as a possible noninformative answer.

Table 1. Comparison of answers for the normal point null test.

x	0	1.68	1.96	2.58
Least favorable Bayesian answer	0.5	0.196	0.128	0.035
Posterior Bayes probability(Aitkin (1991))	0.5	0.256	0.172	0.048
Noninformative answer	0.285	0.089	0.055	0.014
p-value	1	0.093	0.05	0.01

Let us turn now to the behavior of the estimator (2.5). First, it is strictly smaller than the lower bound (2.7) of the Bayesian estimators obtained by Berger and Sellke (1987). Again, it may seem paradoxical that the noninformative answer does not belong to the range of the Bayesian answers but, contrary to point estimation, testing settings allow for discontinuities between proper and improper priors. Moreover, the bound (2.7) was obtained for $\varrho_0 = 1/2$, while ϱ_0 depends on σ^2 in our case. The difference between (2.5) and (2.7) also shows that, although (2.7) appears as the *least favorable Bayesian answer*, it still corresponds to an informative setting and, therefore, that the use of an *informative* (i.e. proper) *prior* makes a significant difference in the answer to a testing problem, even though (2.7) is formally identical to the posterior probability associated with the improper prior

$$\Pi_0(\theta) + \sqrt{2\pi}\lambda(\theta),$$

where $\lambda(\theta)$ denotes the Lebesgue measure on $\mathbf{R}$. This feature definitely separates testing from usual estimation problems but does not necessarily imply that improper priors should not be used.

Table 1 provides some numerical values of the noninformative estimator (2.5) for some values of x. In addition to the above mentioned discrepancy with the least favorable answer, an interesting feature of Table 1 is the closeness of (2.5) and the p-value, $p(x)$, when x is large. Indeed, when the p-value is between 0.10 and 0.01, (2.5) produces essentially the same numerical values. In other words, for the range of x's for which the exact value of $p(x)$ really matters, the noninformative approach leads to the same *decision* as the p-value. (Actually, H_0 will usually be accepted for an answer larger than 0.10 and rejected for an answer smaller than 0.01.) Therefore, *decisionwise*, the two approaches are somehow equivalent.

Obviously, this equivalence does not "rehabilitate" the p-value since the numerous undesirable features pointed out in the previously mentioned papers still exist and a noninformative answer is not necessarily a "good" answer. On the contrary, we could argue that the similarity we have exhibited in this paper rather points out the need for additional (prior) information. Moreover, the closeness of (2.5) and $p(x)$ only occurs on a small (although crucial) range of values of x, and (2.5) is admissible under squared error loss, while $p(x)$ is not (see Hwang et al. (1992)). However, it may also explain why the p-value has survived for such a long period despite its multiple drawbacks. The coincidence of the classical answer with a noninformative answer actually holds in other settings, as shown by Robert and Caron (1991) (who also consider an alternative noninformative approach leading to the same conclusion).

Acknowledgement

This research was supported by NSF Grant DMS-8717799, while the author was visiting Purdue University. I wish to thank J. Berger, G. Casella and E. George for helpful discussions.

References

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *J. Roy. Statist. Soc. Ser.B* **53**, 111–142.

Berger, J. O. (1991). Personal communication.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317–352.

Berger, J. O. and Jeffreys, W. (1991). The application of robust Bayesian analysis to hypothesis testing and Occam's razor. *Proc. Italian Statist. Soc.* **2**, 1–5.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of *P*-values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82**, 112–139.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. In *Bayesian Statistics* (Edited by J. Bernardo, M. DeGroot, D. Lindley and A. Smith). University Press, Valencia.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Amer. Statist. Assoc.* **82**, 106–111.

DeGroot, M. H. (1982). Comment on Shafer (1982). *J. Amer. Statist. Assoc.* **77**, 336–338.

Hwang, J. T., Casella, G., Robert, C. P., Wells, M. and Farrell, R. (1992). Estimation of accuracy in testing. *Ann. Statist.* **20**, 490–509.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edition. Oxford University Press.

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

Lindley, D. V. (1991). Discussion of the paper by Aitkin. *J. Roy. Statist. Soc. Ser.B.* **53**, 130–131.

Robert, C. P. and Caron, N. (1991). Noninformative Bayesian testing and neutral Bayes factors. Technical report #148, L.S.T.A., Université Paris 6.

Shafer, G. (1982). Lindley's Paradox (with discussion). *J. Amer. Statist. Assoc.* **77**, 325–334.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser.B* **42**, 213–220.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. Ser.B* **44**, 377–387.

URA CNRS 1378, Faculté des Sciences, Université de Rouen, 76821 Mont Saint Aignan, France.

(Received November 1991; accepted January 1993)