Comment

Arnold Zellner

1. INTRODUCTION

In this stimulating and important paper, Bayes factors, posterior probabilities and P-values are considered in relation to the problem of using data to evaluate precise or sharp null hypotheses, for example $\theta = 0$ or $\theta = 1.0$. This is a very basic problem encountered in all areas of science and thus the fact that the authors, along with Jeffreys (1967) and others, conclude that widely used P-values are unsatisfactory is noteworthy. This conclusion has important implications not only for textbook treatments of the theory of testing, but also for applied scientific work.

The authors explain Jeffreys' approach to testing and show that P-values diverge markedly from posterior probabilities associated with sharp null hypotheses. They also derive lower bounds for Bayes factors and posterior probabilities and provide some advice in answer to the question, "What should be done?" Although some of the points that the authors raise have appeared in the literature, it is doubtful that they have been expressed as clearly and forcefully as in the present paper. However, as might be expected in such a controversial area, where are some points that deserve further discussion. See, e.g., Jeffreys (1967, Chapters V to VII), Edwards, Lindman and Savage (1963), Jaynes (1984) and Zellner (1971, 1980, 1984) for earlier considerations of testing issues and computation of Bayes factors for a number of problems. After taking up some general points, I shall turn to technical points and then provide some concluding remarks.

2. GENERAL POINTS

Point 1. H_1: $\theta = \theta_0$ *versus* H_2: $|\theta - \theta_0| < \varepsilon$, $\varepsilon > 0$, *Given.* For years I have stated that we should be able to test either H_1 or H_2 or both. To say that H_2 is "realistic" or "true" is to make an unwarranted general *a priori* statement about the "real world." To be scientific, one can compute a Bayes factor for H_1 versus H_2, as Jeffreys (1967, page 367) suggests. He also states, "I think, however, that it is both impossible and undesirable [to replace H_1 with H_2]" (page 367). I won't review Jeffreys's arguments here since they are readily available. It does seem relevant to remark that in $s = .5gt^2$ and in $E = mc^2$, the powers of t and c are predicted by physical theory to be exactly equal to 2. Also the coefficient of t^2 in the former relation is exactly $.5g$ and of c^2 in the latter exactly m. Many other examples of sharp or precise hypotheses can be given and thus it is incorrect to exclude such hypotheses *a priori* or term them "unrealistic" and important to be able to test them well as Berger and Delampady indicate.

Point 2. *Laplace's versus Jeffreys' Approaches to Testing.* Jaynes (1980) raised this point, which is equivalent to Rejoinder 3: Just Use Confidence Intervals of Berger and Delampady. As pointed out in Jeffreys (1967), Zellner (1971, 1980) and Berger and Delampady, the background prior information is different when there is a suggested value θ_0 for θ. However, in Zellner and Siow (1979) and Zellner (1984) it is shown that consideration of three hypotheses, H_1: $\theta = 0$, H_2: $\theta > 0$, and H_3: $\theta < 0$ with prior probabilities, π_1, π_2 and π_3, respectively, and truncated Jeffreys' Cauchy priors under H_2 and H_3 leads to a Bayes factor for H_2 versus H_3 that can be exactly equal to the Laplacian or diffuse prior credible region results. Also, consideration of all three hypotheses together yields a synthesis of the considerations in the present paper and those in Cassella and Berger (1987).

Point 3. Berger and Delampady err in calling Jeffreys', my and some others' Bayesian testing procedures "mechanical" or "automatic" or "default" or "conventional" or "objective." Jeffreys (1967, page 252) explains that in testing there may be very little previous information or a great deal. If there is a great deal of prior information, Jeffreys (1967, page 252) and others would use an appropriate prior distribution to represent it. Although Jeffreys mainly analyzed the situation of "little previous information" in his book, this does not imply at all that he would use these procedures when there is a great deal of previous information. I have expressed similar views (Zellner, 1980, 1984). However, it is useful, as I believe that Berger and Delampady recognize, to have testing results for the case of little previous information "on the shelf" to be used *when appropriate.* Further, the priors that Jeffreys used for the normal mean problem and others can be given different location and scale parameters without much difficulty as Jeffreys (1967),

Arnold Zellner is H. G. B. Alexander Distinguished Service Professor of Economics and Statistics, Graduate School of Business, University of Chicago, Chicago, Illinois 60637.

Berger and Delampady and others have recognized and thus can represent a fairly broad range of previous information.

Point 4. *Do Classical and Likelihood or Bayesian Answers Typically Agree?* This is a difficult question to answer because it's difficult to know what is meant by "classical answers." For example, in his classic work on hypothesis testing Lehmann (1959) states:

> "Another consideration that frequently enters into the specification of a significance level is the attitude toward the hypothesis before the experiment is performed. If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low" (page 62).

If one follows Lehmann's advice, he would not always reject a null hypothesis, say $\theta = \theta_0$, when $t = 1.96$, but would use a larger critical value for t if he firmly believes in the null hypothesis. Also, his probability on the null would not be $\pi_0 = .5$, but would probably be higher. Thus, the analysis of Berger and Delampady does not directly apply to a "Lehmann tester." However, it is doubtful that most testers are "Lehmann testers."

Further, as noted in Zellner (1971, page 304), most non-Bayesians would adjust a significance level from say .05 to .03 and to lower values as the sample size grows, perhaps in order to balance probabilities of errors of the first and second kind. With such an adjustment, the Jeffreys-Lindley paradox disappears, as is perhaps now well-known. Still, it is difficult to know how to adjust the significance level as the sample size grows. Also, I have the impression from my own experience, from Jeffreys' report of what astronomers do and from talking with others that many tend not to reject a null hypothesis when $t = 1.96$, but view the matter as a situation in which more information is needed. Finally, I don't believe that Berger and Delampady have pinpointed why many, at least in business and economics use P-values. I believe that it is because they find it very difficult to select an appropriate significance level. Thus, many report P-values in their papers and leave it to the reader to interpret them.

Point 5. *No Alternative Hypothesis?* This question is discussed in Zellner (1980, 1984) where it is pointed out that without an alternative hypothesis, it is difficult to select an appropriate test statistic. That is, as Jeffreys (1967, page 385) points out, there is always some function of the data that looks unusual given a null hypothesis. Usually, selection of a test statistic implies consideration of a specific type of departure from the null. For example, if my null hypothesis is that the dollar-yen exchange rate follows a Gaussian random walk with a zero drift parameter, there are obvious alternative hypotheses available. Thus, frequently encountered null hypotheses such as "no effect" or "all variation is random until shown otherwise" are usually accompanied by almost obvious alternative hypotheses as in the case of the dollar-yen exchange rate example above. If one cannot think of a relevant, important alternative to a null hypothesis, perhaps he should not be considering hypotheses at all. Finally, as Jeffreys (1967) asks, "Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place?" (page 390).

3. TECHNICAL POINTS

Point 1. Instead of devoting a huge amount of effort to approximating a sharp or precise null hypothesis by an appropriate null hypothesis, which is a misguided effort in my opinion, I recommend considering the following hypotheses and getting Bayes factors and posterior probabilities for them: H_1: $\theta = \theta_0$, H_2: $0 < \theta - \theta_0 < \varepsilon$, H_3: $-\varepsilon < \theta - \theta_0 < 0$, H_4: $\theta - \theta_0 > \varepsilon$ and H_5: $\theta - \theta_0 < -\varepsilon$, where $\varepsilon > 0$ is given. Also, Bayes factors can be computed for H_1 versus various unions of the remaining hypotheses.

Point 2. Jeffreys (1967, page 274) has analyzed the normal mean problem with a known value for σ using a Cauchy prior under the alternative hypothesis as well as in the case in which σ's value is unknown. For $\theta = 0$ versus $\theta \neq 0$ with σ known, he obtains a Bayes factor $\doteq \sqrt{\pi n/2}(1 + t^2/n)\exp\{-t^2/2\}$, which can be compared with Berger and Delampady's Bayes factor $= (1 + n)^{1/2}\exp\{-t^2/2(1 + 1/n)\}$, used to compute entries in Table 1 of Berger and Delampady. Although the Bayes factors of Jeffreys and Berger and Delampady are qualitatively similar in their behavior, they do provide different values for Bayes factors. For example, when $t = 1.960$ and $n = 100$, Jeffreys's Bayes factor = 1.91, whereas Berger and Delampady's Bayes factor = 1.50. Jeffreys's reasons for choosing the particular Cauchy prior that he employed, and not, e.g., a normal prior, are not as inconsequential as Berger and Delampady apparently suggest. See Jeffreys (1967, page 268–270 and page 273) for discussion of this point. Further, he obtains his Cauchy prior by placing a uniform prior on a function of the Jeffreys-Kullback-Leibler information divergence measure for the data densities under the null and alternative hypotheses (Jeffreys, 1967, page 275). This rule for generating priors and considerations in Jeffreys (1967, page 268) indicate why θ and σ are not viewed independent *a priori* and also appears to rationalize the view that when there is little prior information, information about θ will be expressed in terms of θ/σ.

Last, the Jeffreys-Kullback-Leibler information divergence approach to generate priors, mentioned above, does generate the same prior for whomever uses the approach and thus may be termed "objective" in this sense.

Point 3. The authors state, perhaps diplomatically, "We would feel comfortable with the use of any of the conventional choices of priors referenced [in Section 2]." This is an unsatisfactory statement in my opinion because use of one of the priors referenced results in a posterior probability for the null hypothesis, $\theta = \theta_0$ *which cannot exceed* .5, an absurd result that is not a property of the results of Jeffreys and others for this problem. Further, as stated above, the reasonable conventional priors that Jeffreys and others put forward are useful when little information is available. Perhaps Berger and Delampady are suggesting that they are also useful when a great deal of previous information is available. In my opinion, this view is erroneous.

Point 4. Is the class of prior densities in (10) reasonable? In connection with the normal mean problem, $\theta = 0$ versus $\theta \neq 0$, with σ's value unknown, Jeffreys (1967, page 269) argued that if $n = 1$, the Bayes factor should be equal to one that requires that the prior density for θ/σ under the alternative must be symmetric and proper. However, the requirement that a very large value of $t = \sqrt{n}\,\bar{x}/s$ leads to a Bayes factor close to zero led him to adopt a Cauchy rather than, for example, a normal density under the alternative. Thus, the requirement of Berger and Delampady that the prior density be nonincreasing in $|\theta - \theta_0|$ seems too weak. Also, in other situations with more information available, it may be wise to consider asymmetric priors, priors not centered at the null hypothesis, etc. Thus, I am suggesting that broader prior densities be selected in reasonable ways for the problem at hand and relevant Bayes factors be calculated as in the case of considering $\theta = 0$, $\theta > 0$ and $\theta < 0$ rather than $\theta = 0$ and $\theta \neq 0$. These considerations make me believe that introduction of the class of priors in (10) and associated bounds will introduce too much arbitrariness in analyses of hypotheses.

4. CONCLUDING REMARKS

The authors have provided us with a significant paper for which we should all be grateful. The main conclusion, forcefully and convincingly demonstrated, that P-values should be abandoned requires immediate, serious attention by all testers. That the authors' conclusion is in agreement with views put forward earlier by others does not detract from its importance but only magnifies it.

Second, I believe that the authors have not adequately described the general approach of Jeffreys and his specific treatment of testing hypotheses about a normal mean.

Third, Berger and Delampady honestly and forthrightly recognize some deficiencies associated with their class of densities in (10) and their bounds. I have suggested that less general, important alternatives are usually available and Bayes factors for them should be computed to get a hold on the sensitivity of results to specific, relevant broader assumptions. For example, in testing the hypothesis of a Gaussian random walk, H_1: $y_t = y_{t-1} + \varepsilon_t$ versus H_2: $y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t$, a second order Gaussian autoregressive process, the prior density for α_1 and α_2 might be centered at (1, 0) if little is known. Alternatively, the prior density might be centered at values of α_1 and α_2 that give rise to oscillatory behavior of the process with a reasonable period if this information about the behavior of the process under H_2 is available. Also, for $y_t = \rho y_{t-1} + \varepsilon_t$ and H_1: $\rho = 1$ versus H_2: $\rho \neq 1$, Mañas-Anton (1986) found it unreasonable to consider values of ρ much larger than 1 under H_2 because these would lead to highly unlikely behavior of y_t under H_2. These examples indicate how available information can be employed to define reasonable, relevant alternative hypotheses. It appears necessary for statisticians testing hypotheses to have a good understanding of subject matter considerations in order to obtain sensible results.

ACKNOWLEDGMENT

This research was financed in part by the National Science Foundation and by income from the H. G. B. Alexander Endowment Fund, Graduate School of Business, University of Chicago.

ADDITIONAL REFERENCES

JAYNES, E. T. (1980). Discussion. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 618–629. University Press, Valencia. Reprinted in *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics* (R. D. Rosenkrantz, ed.) 378–400. Reidel, Dordrecht.

JAYNES, E. T. (1984). The intuitive inadequacy of classical statistics. *Epistemologia* **7** (*Special Issue on Probability, Statistics and Inductive Logic*) 43–74.

JEFFREYS, H. (1967). *Theory of Probability*, 3rd rev. ed. Oxford Univ. Press, London.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

MAÑAS-ANTON, L. A. (1986). Empirical regularities in short-run exchange rate behavior. Ph.D. dissertation, Dept. Economics, Univ. Chicago.

ZELLNER, A. (1980). Reply. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 638–643. University Press, Valencia. A version with typographical errors corrected is available from the author on request.

ZELLNER, A. and SIOW, A. (1979). On posterior odds ratios for sharp null hypotheses and one-sided alternatives. Technical Report, H. G. B. Alexander Research Foundation, Graduate School of Business, Univ. Chicago.