Comment

M. J. Bayarri

I would like first of all to congratulate the authors for such an interesting and enjoyable paper. The widespread use of tests of hypotheses to statistically analyze experimental data, together with the failure of classical methods to make statements about the truth of H_0 in a given problem, has almost unavoidably resulted in interpreting P-values as a measure of the evidence against H_0 provided by the data at hand. The authors show how misleading this procedure can be. They also provide the statistical community with some "automatic" tools as easy to implement as P-values and with a better performance, but with the remarkable suggestion of not just substituting a routine statistical analysis (P-values) by another one.

Berger and Delampady also justify the habitual practice of testing a point null hypothesis by showing that a point null can be a reasonable approximation of a precise interval null, and the conditions under which this approximation is appropriate. It is at this point that I would like to raise a complaint more than a real disagreement. It seems to me that their treatment is unfair to statisticians who use conditional measures of evidence against H_0, for it rules out a lot of interesting situations. I shall make my point clearer. In the examples of Section 2, all that frequentist statisticians have to care about in approximating an interval null by a point null without much error, is the length of their interval null being suitably small compared with the sample standard deviation. (By the way, this care would prevent them from using the testing of a point null when n is very large.) On the other hand, Bayesians who want to approximate the same precise hypothesis (11) find that they have to care not only about that in a similar way as the frequentists, but also ought to "have in mind a prior density, $\pi(\theta)$, which is continuous but sharply spiked near θ_0." Although I don't deny the fact that these sharply spiked densities often represent the prior beliefs of statisticians performing tests of precise hypotheses, I do claim that this is not the case in many interesting situations.

For instance, if the prior density, $\pi(\theta)$, belongs to a conjugate class of prior distributions for any of the common models, then no matter how concentrated $\pi(\theta)$ is around θ_0 it would not usually be possible to approximate a precise null by a point null. Another interesting class of problems in which Bayesian statisticians cannot use this approximation is that of testing precise null hypotheses that are judged to be false *a priori*. This situation is quite frequent. As a matter of fact, the scientific literature is overwhelmed with "significant results," which are a natural consequence of the misuse of statistical methods in many areas of application through the almost exclusive reliance on tests of hypotheses (Zellner, 1980; DeGroot and Mezzich, 1985). Some of these significant results can be due to the unfavorable treatment given to the null hypothesis by the P-value (as clearly shown in this paper) even if the scientist believed it to be true *a priori*, but confronted with this overpresence of significant results, it is natural to suspect that some of these tests of hypotheses have been carried out with the sole purpose of rejecting the null hypothesis. Under such an assumption, a Bayesian can no longer have at her or his disposal the useful approximations described in Section 2.2.

But maybe the most interesting problems to which the methods described in this paper cannot be applied (as the authors explicitly recognize in Section 5) are those of goodness-of-fit tests when there is not a spiked concentration of prior beliefs around the model in the way described in Section 2. Checking models usually is (or should be) a preliminary step in every parametric statistical analysis. But models are seldom thought of as *true*, they are just simplifications to explain the random behavior of some quantities. Also, this is one of those statistical problems in which a statistician could wish to "let the data speak for themselves," that is, to use an "objective" or "reference" prior, perhaps in addition to her or his prior distribution.

For these situations I should like to discuss a method to carry out the testing of a precise null that still preserves the special nature of θ_0. In a goodness-of-fit scenario, the hypothetical model is special to us because it is *useful* for us to use this particular model instead of a more complicated one, usually because statistical techniques are well developed and studied for this particular model, or because it is the one implemented in the statistical computer packages at our disposal. Accordingly, we will make θ_0 "special" to us in terms of the utility function instead of in terms of the prior distribution.

M. J. Bayarri is Titular Professor in the Department of Statistics and Operations Research, Faculty of Mathematics, University of Valencia, Av. Dr. Moliner 50, Burjasot 46100 Valencia, Spain.