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.

Consider the situation described in Example 5 in which it is desired to test

$$(1) \qquad H_0\colon F = F_0 \quad \text{versus} \quad H_1\colon F \neq F_0$$

where F_0 is a specified distribution. If we transform the data by F_0 then testing (1) is equivalent to testing whether or not the transformed data can be assumed to be a random sample from a uniform distribution. Testing for uniformity can be carried out in several ways. One interesting possibility is to choose a parametric family of distributions $\{h(x \mid \theta),\ \theta \in \theta\}$ for $0 < x < 1$ such that it contains the uniform distribution for a particular value θ_0, say, of θ. In this way, testing (1) is reduced to testing

$$(2) \qquad H_0\colon \theta = \theta_0 \text{ versus } H_1\colon \theta \neq \theta_0,$$

but it should be kept in mind that in order for $h(x \mid \theta)$ to be useful it should be rich enough to model interesting alternatives and "neighborhoods" of F_0 and also be reasonably easy to handle. In addition, the parameter (or parameter vector) indexing the family should have an intuitive meaning, not only to ease the task of assessing prior distributions, but also and most importantly to be able to draw conclusions about the shape of the "true" distribution should H_0 be rejected.

In testing (2), we will adopt a decision-oriented approach and follow the method proposed by Ferrandiz (1985). Call d_0 the decision of continuing with the statistical analysis of data as if F_0 were the "true" model, and d_1 the decision of rejecting F_0 as a sensible explanation of the data. After the previous steps mentioned above, d_0 can be reformulated as accepting H_0 in (2) and d_1 as accepting H_1. For this type of decision problem it seems natural to assume that the increase in utility derived from choosing d_1 (reject H_0) is a continuous, nondecreasing function v of some "distance" δ between the true parameter value θ and the hypothetical one θ_0. We use quotation marks around the term distance because we are not going to require δ to be a proper distance, but only a kind of discrepancy measure between $h(x \mid \theta)$ and $h(x \mid \theta_0)$. Thus we assume

$$u(d_1, \theta) - u(d_0, \theta) = v\{\delta(\theta, \theta_0)\},$$

where $u(d_i, \theta)$ is the utility of choosing d_i when θ is the value of the parameter.

Now, from the Bayesian approach, d_1 is the optimal decision if and only if

$$(3) \qquad E\{v(\delta) \mid \mathbf{x}\} > 0,$$

where expectation is taken with respect to the posterior distribution of θ. If in particular we take v to be linear,

$$(4) \qquad v(\delta) = a\delta + b \quad (a > 0),$$

then condition (3) reduces to $E\{\delta(\theta, \theta_0) \mid \mathbf{x}\} > \delta_0$, where δ_0 is a constant. Thus, the null hypothesis is to be rejected whenever the function of the (transformed) data

$$U(\mathbf{x}) = E\{\delta(\theta, \theta_0) \mid \mathbf{x}\}$$

exceeds some preassigned value δ_0.

The Bayesian decision rule so exposed looks like the classical one, but note that δ_0 is part of the utility function and has to be interpreted in a special way. As a matter of fact, if we take $a = 1$ in (4) and naturally require $\delta(\theta_0, \theta_0) = 0$, then $\delta_0 = -b = u(d_0, \theta_0) - u(d_1, \theta_0)$ measures (in utility terms) the relative advantage of using the simpler model when it is adequate. Alternatively, we can write $u(d_1, \theta_0) = u(d_0, \theta_0) - \delta_0$ and so interpret δ_0 as a penalty for using a more elaborate model when a simpler one would suffice.

This approach to goodness-of-fit problems was proposed by Bayarri (1985), where a particular family $h(x \mid \theta)$ was selected and justified, and different δ as well as various methods for selecting δ_0 were studied. It was there found to be particularly appealing in goodness-of-fit problems to choose δ as the Kullback-Leibler directed divergence between $h(x \mid \theta)$ and $h(x \mid \theta_0)$, but other δ's such as quadratic or absolute error were also considered. It was also shown that keeping the same level of significance α as the sample size increases can be interpreted as changing the utility function with the sample size, namely, making δ_0 progressively smaller. Thus, keeping α fixed is equivalent to saying that when F_0 is adequate, the bigger the sample size, n, the less useful it is for us to use F_0, and this does not seem very realistic.

The original formulation of Ferrandiz (1985) was applied to testing the mean vector in a multivariate normal distribution; δ was selected to be the Mahalanobis distance and δ_0 the $1 - \alpha$ quantile of the sampling distribution of $U = U(\mathbf{X})$ under the null hypothesis. With this particular choice, and with an "objective" prior for the parameter vector, he reproduced the standard frequentist results, and by establishing a one to one transformation between δ_0 and the level of significance α he showed how to modify the latter as the sample size increases in order to be coherent.

Of course this method can also be applied to the testing situations described in the paper. Take for instance Example 1 in which it is desired to test

$$H_0\colon \theta = \theta_0 \quad \text{versus} \quad H_1\colon \theta \neq \theta_0.$$

If it is thought that θ_0 is special because of the utility it has for us to take $\theta = \theta_0$ when θ is close to θ_0, then we can assume that

$$u(d_1, \theta) - u(d_0, \theta) = \delta(\theta, \theta_0) + b,$$

where d_1 means rejecting H_0 and b is a constant. An appropriate distance δ between θ and θ_0 in this case is the standardized square distance (Mahalanobis distance)

$$\delta(\theta, \theta_0) = \{(\theta - \theta_0)/\sigma\}^2,$$

which happens to be twice the Kullback-Leibler divergence between the $N(\theta, \sigma^2)$ and the $N(\theta_0, \sigma^2)$ distributions. According to the discussion above, we will reject H_0 if and only if

$$E[\delta(\theta, \theta_0) \mid \mathbf{x}] > \delta_0.$$

If we take the usual "objective" prior for this problem, $\pi(\theta) \propto 1$, then the posterior distribution of θ is simply $N(\bar{x}, \sigma^2/n)$ so that

$$U(\mathbf{x}) = E[\delta(\theta, \theta_0) \mid \mathbf{x}]$$
$$= (1/n) + (\bar{x} - \theta_0)^2/\sigma^2 = (1 + T^2)/n$$

where T is given in Example 1. Then we will reject H_0 whenever $T^2 > c(n) = n\delta_0 - 1$.

We could explicitly seek an analogy with the classical methodology and thus select δ_0 to be the $1 - \alpha$ quantile of the sampling distribution of $U = U(\mathbf{X})$ under the null hypothesis, where α is the level of significance (not the P-value as in Example 1). In this case, with this *particular* value of n, we would reproduce the frequentist test procedure. But if the value of n changes, δ_0 still must have the same value, so that $c(n)$ must change. Thus, the frequentist rule of choosing $c(n)$ so that the test has size α can have a Bayesian interpretation as long as α changes accordingly with the results above. Of course, this example is just a particular case of the problem studied in Ferrandiz (1985).

ADDITIONAL REFERENCES

BAYARRI, M. J. (1985). A Bayesian test for goodness-of-fit. Technical Report, Departamento de Estadistica e Investigación Operativa, Univ. Valencia.

DEGROOT, M. H. and MEZZICH, J. E. (1985). Psychiatric statistics. In *A Celebration of Stastistics: The ISI Centenary Volume* (A. C. Atkinson and S. E. Fienberg, eds.) 145–165. Springer, New York.

FERRANDIZ, J. R. (1985). Bayesian inference on Mahalanobis distance: An alternative approach to Bayesian model testing. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 645–654. North-Holland, Amsterdam.

ZELLNER, A. (1980). Statistical analysis of hypotheses in economics and econometrics. *Proc. Amer. Statist. Assoc. Bus. Econ. Statist. Sec.* 199–203.

Comment

George Casella and Roger L. Berger

We congratulate Berger and Delampady on an informative paper. However, we do not believe that the point null testing problem they have considered reflects the common usage of point null tests. Their main thesis is that the frequentist P-value overstates the evidence against the null hypothesis although the Bayesian posterior probability of the null hypothesis is a more sensible measure. A second point of their paper is that point null hypotheses are reasonable approximations for some small interval nulls. We disagree with both of these points.

The large posterior probability of H_0 that Berger and Delampady compute is a result of the large prior probability they assign to H_0, a prior probability that is much larger than is reasonable for most problems in which point null tests are used. Replacing a large prior probability for a point by an equally large prior probability for a small interval about the point does not remedy the problem. It only replaces one unrealistic problem with another. We will argue that given a reasonably small prior probability for an interval about the point null, the posterior probability and the P-value do not disagree. Before moving to the main points of our rejoinder, however, we would like to make a general comment.

Contrary to what Berger and Delampady would have us believe, a great many practitioners should not be testing point nulls, but should be setting up confidence intervals. Interval estimation is, in our opinion, superior to point null hypothesis testing, Rejoinder 3 of Berger and Delampady notwithstanding. However, we will not argue about the appropriateness of the test of a point null. Instead, we will argue the following: Given the common problems in which point null tests are used, the Bayesian measure of evidence, as exemplified by equation (4) of Berger and Delampady is not a meaningful measure. In fact, it is not the case that P-values are too small, but rather that Bayes point null posterior probabilities are much too big!

George Casella is Associate Professor, Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, New York 14853. Roger L. Berger is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695.