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 Estadística 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.