because a test contrast has been standardized by a null hypothesis standard error. Such a practice may be computationally convenient, as with score tests, but its negative features should not be overlooked.

One must agree that the operational interpretation of P values must be made relative to the amount of information available in the data, as expressed through ancillary statistics. Barnard (1982) argued cogently for this in the context of repeated significance tests, where a fixed cutoff for P values can lead to drastic loss of overall power. Of course confidence statements automatically account for available information, if proper conditioning is employed.

ADDITIONAL REFERENCES

Barnard, G. A. (1982), "Conditionality Versus Similarity in the Analysis of 2 × 2 Tables," in *Statistics and Probability: Essays in Honor of C. R. Rao,* eds. G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, Amsterdam: North-Holland, pp. 59–65.

Racine, A., Grieve, A. P., Fluhler, H., and Smith, A. F. M. (1986), "Bayesian Methods in Practice: Experiences in the Pharmaceutical Industry" (with discussion), *Applied Statistics*, 35.

Comment

JAMES M. DICKEY*

What should our reaction be to the results announced in these two articles? What do they actually say to us, and what difference should it make in statistical practice? Before attempting to answer these questions, I would like to bring up a few relevant points.

Example 1, which runs through the Berger–Sellke article, is introduced by using the normal distribution, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, as the conditional prior uncertainty given the alternative H_1. This distribution has the same variance as the sampling process. Consider, however, the generalization to an arbitrary prior variance, $\theta \sim \mathcal{N}(\theta_0, \tau^2)$, say $\tau^2 = \sigma^2/n^*$. In this notation, n/n^* represents the ratio $\tau^2/(\sigma^2/n)$ of the prior variance to the sampling variance of the sample mean. Unless I am mistaken, the expressions and tables in Sections 1 and 2 for the posterior probability $\Pr(H_0 \mid x)$ hold again for the more general case by merely replacing the variable n by n/n^* throughout. (The variable t retains its original definition in terms of the sample size n.) In many, if not most, areas of application, the conditional prior variance τ^2 is typically larger than the sampling variance σ^2. So the ratio n/n^* is larger than n, and one would find oneself looking further over in the right-hand (large-n) direction in Table 1 than if one pretended one's τ^2 equaled σ^2. In such applications, the effect touted here by Berger and Sellke is strengthened. The posterior probability of the null hypothesis tends not to be as small as the P value of the traditional test.

Theorems 2, 4, and 7 give lower bounds for the posterior probability of the null hypothesis in the case in which the corresponding prior probability π_0 is equal to $\frac{1}{2}$. Of course, the Bayes factor B, the ratio of posterior odds for H_0 to the corresponding prior odds $\pi_0/(1 - \pi_0)$, does not depend on π_0. Hence one is tempted to ask for versions of these theorems stated in terms of the Bayes factor. It is curious to see that the limits claimed for large t in these theorems do not appear in the accompanying tables as visible tendencies for increasing t. Rather, an opposite tendency, to move away from the limit, is exhibited. So it would seem that the limits are meaningless except for exorbitantly large values of t. (That is, meaningless in practice: H_0 would be strongly rejected by all methods before the limit would have any effect?) Have the authors done any investigating to see where the limits begin to take effect?

To my mind, the Casella–Berger article further supports the thesis of Berger and Sellke. Theorems 3.2 and 3.3 of Casella and Berger concern an infimum over a class of prior distributions. So the smallest corresponding posterior probability of one-sided H_0 equals the traditional P value, and this equality is attained for the extreme constant prior pseudodensity. That is, reasonable prior distributions give posterior probabilities for H_0 that are larger than the traditional P value, though perhaps not as much larger as in the case of a point null hypothesis.

By the way, the constant prior pseudodensity appears here in the second of its two legitimate roles in inference, as follows. Bayesian scientific reporting requires a report of the effect of the observed data on a whole range of prior distributions, keyed to context-meaningful prior uncertainties (Dickey 1973). "Noninformative" prior pseudodensities are sometimes useful for such reporting in two ways:

1. Such a prior can serve as a device to give a simple posterior distribution that approximates the posterior distributions from prior probability distributions expressing relevant context uncertainties. This approximation is quantified by L. J. Savage's "stable estimation" or "precise measurement" (Edwards, Lindman, and Savage 1963; Dickey 1976).
2. Such a prior can serve as a device to give bounds on posterior probabilities over classes of context-relevant prior distributions.

* James M. Dickey is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455. This work was supported by National Science Foundation Research Grant DMS-8614793.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

What should our attitude now be concerning P values? Berger and Sellke note that nonstatisticians tend to confuse the P value and the posterior probability of the null hypothesis. As pointed out in Good (1984), even the most respected statisticians can make the same mistake. The present works reinforce the distinction between sampling probability and posterior probability.

It has long seemed to me that the P value reports an interesting fact about the data. I once speculated to Dennis Lindley that the P value might offer a quicker and cruder form of inference than the Bayes factor. He replied by asking whether what I meant was analogous to comparing an orchestra with a tom-tom.

ADDITIONAL REFERENCES

Dickey, James M. (1976), "Approximate Posterior Distributions," *Journal of the American Statistical Association,* 71, 680–689.

Good, I. J. (1984), "An Error by Neyman Noticed by Dickey" (C209), in "Comments, Conjectures, and Conclusions," *Journal of Statistical Computation and Simulation,* 20, 159–160.

Comment

STEPHEN B. VARDEMAN*

Berger, Sellke, Casella, and Berger deserve our thanks for a most readable and thorough accounting of the problem of comparing p values and posterior probabilities of H_0. They have laid out in very clear fashion the history of the problem, a full array of technical points, and their arguments from the technical points to general conclusions. Their articles should help all of us, card-carrying Bayesians, militant frequentists, and fence-sitters like myself, to sort this issue out to our own satisfaction.

My view from the fence is that in spite of the fact that the articles are well done, there is nothing here very surprising or that carries deep philosophical implications. We all know that Bayesian and frequentist conclusions sometimes agree and sometimes do not, depending on the specifics of a problem. These articles seem to me to reinforce this truism. For example, I read the Casella/Berger Theorem 3.4, the argument behind it, and their subsequent discussion as confirmation that essentially anything can be possible for a posterior probability for H_0, *depending on how one is allowed to move prior mass around on* H_0 *and* H_1. (Of course, the simplest demonstration that nearly anything can be possible can be made by using arbitrary two-point priors in a composite versus composite case.)

Whether or not a Bayesian analysis can produce a small posterior probability for H_0 is largely a function of whether or not (staying within whatever rules are imposed by the problem structure and restrictions adopted for the prior) one can move the prior mass on H_0 "away from the data," at least as compared with the location of the prior mass on H_1. If this can be done, the posterior probability of H_0 can be made small, otherwise it cannot.

Take, for example, the Jeffreys–Lindley "paradox" discussed by Berger and Sellke. To maintain a p value that is constant with n (i.e., a constant value of t), one must send $\overline{X}_n$ (the data) to θ_0. The nonzero mass on H_0 is trapped at θ_0, while the mass on H_1 is all passed by as $\overline{X}_n \to \theta_0$. Why should anyone then be surprised that the posterior probability assigned to H_0 tends to 1?

Moving to a different point, I must say that I find the "spike at θ_0" feature of the priors used by Berger and Sellke and many before them to be completely unappealing. In fact, contrary to the exposition of Berger and Sellke, I think that the appeal of such priors *decreases* with increasing π_0. Unlike that of Casella and Berger, my objection has nothing to do with "impartiality" (indeed I question whether such a concept can have any real meaning), but is of a more elementary nature. The issue is simply that I do not believe that any scientist, when asked to sketch a distribution describing his belief about a physical constant like the speed of light, would produce anything like the priors used by Berger and Sellke. A unimodal distribution symmetric about the current best value? Probably. But with a spike or "extra" mass concentrated at θ_0? No.

Competent scientists do not believe their own models or theories, but rather treat them as convenient fictions. A small (or even 0) prior probability that the current theory is true is not just a device to make posterior probabilities as small as p values, it is the way good scientists think! The issue to a scientist is not whether a model is true, but rather whether there is another whose predictive power is enough better to justify movement from today's fiction to a new one. Scientific reluctance to change theories is appropriately quantified in terms of a cost structure, not by concentrating prior mass on H_0. In this regard, note that although the "spike at θ_0" priors are necessary to produce nontrivial Bayes rules (i.e., ones that sometimes "accept") for a zero–one type loss structure in the two-sided problem, other competing cost structures do not require them for a Bayesian formulation of the testing

* Stephen B. Vardeman is Professor, Statistics Department and Industrial Engineering Department, Iowa State University, Ames, IA 50011.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods