Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence: Comment
Author(s): David V. Hinkley
Source: *Journal of the American Statistical Association*, Vol. 82, No. 397 (Mar., 1987), pp. 128-129
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2289134
Accessed: 28-03-2015 14:33 UTC

are included. To aid in making the bibliography more complete I exercise the rights of a senior citizen and list 28 additional relevant publications of which I have read every word (10 of them are in the conscientious reference list of B&S): (a) items C73, C140, C144, C199, C200, C201, C209, C213, C214, and C217 in *Journal of Statistical Computation and Simulation* (1984); (b) Items 13 (pp. 91–96), 82, 127 (pp. 127–128), 174, 398 (p. 35), 416, 547, 603B (p. 61), 862, 1234 (pp. 140–143), 1278 (regarding Bernardo), 1320–C73, 1396 (pp. 342–343), 1444, and 1475–C144 in the bibliography (pp. 251–266) in Good (1983); (c) Good (1955/1956, p. 13; 1981; 1983, indexes; 1986; in press a,b). To these may be added the thesis of my student Rogers (1974) and a further reference relevant to C&B, Thatcher (1964).

ADDITIONAL REFERENCES

Good, I. J. (1955/1956), Discussion of "Chance and Control: Some Implications of Randomization," by G. S. Brown, in *Information Theory, Third London Symposium 1955,* London: Butterworth's, pp. 13–14.

——— (1957), "Saddle-Point Methods for the Multinomial Distribution," *Annals of Mathematical Statistics,* 28, 861–881.

——— (1976), "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *The Annals of Statistics,* 4, 1159–1189.

——— (1981), Discussion of "Posterior Odds Ratio for Selected Regression Hypotheses," by A. Zellner and A. Siow, *Trabajos de Estadistica y de Investigacion Operativa,* 32, No. 3, 149–150.

——— (1982a), Comment on "Lindley's Paradox," by Glenn Shafer, *Journal of the American Statistical Association,* 77, 342–344.

——— (1982b), "Standardized Tail-Area Probabilities" (C140), *Journal of Statistical Computation and Simulation,* 16, 65–66.

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

——— (1984b), "A Sharpening of the Harmonic-Mean Rule of Thumb for Combining Tests 'in Parallel' " (C213), *Journal of Statistical Computation and Simulation,* 20, 173–176.

——— (in press a), "A Flexible Bayesian Model for Comparing Two Treatments," C272, *Journal of Statistical Computation and Simulation,* 26.

——— (in press b), "Scientific Method and Statistics," in *Encyclopedia of Statistical Science* (Vol. 8), eds. S. Kotz and N. L. Johnson, New York: John Wiley.

Good, I. J., and Crook, J. F. (1974), "The Bayes/Non-Bayes Compromise and the Multinomial Distribution," *Journal of the American Statistical Association,* 69, 711–720.

Jeffreys, H. (1939), *Theory of Probability* (1st ed.), Oxford, U.K.: Clarendon Press.

Rogers, J. M. (1974), "Some Examples of Compromises Between Bayesian and Non-Bayesian Statistical Methods," unpublished doctoral thesis, Virginia Polytechnic Institute and State University, Dept. of Statistics.

Thatcher, A. R. (1964), "Relationships Between Bayesian and Confidence Limits for Predictions" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 26, 176–192.

Comment

DAVID V. HINKLEY*

The authors have added an impressive array of technical results to the main body of work on this subject by Jeffreys, Lindley, and others. The sense of surprise in the first article suggests that statistical education is not as eclectic as one might wish. In my brief comments I should like to mention some of the general issues that should be considered in any broad discussion of significance tests.

First, the interpretation of P value as an error rate is unambiguously objective and does not in any way reflect the prior credibility of the null hypothesis. Rules of thumb aimed at calibrating P values to make them work like posterior probabilities cannot reflect the broad range of practical possibilities: in many situations the null hypothesis will be thought not to be true.

One area where null hypotheses have quite high prior probabilities is model checking, including both goodness-of-fit testing and diagnostic testing. Here specific alternative hypotheses may not be well formulated, and significance test P values provide one convenient way to put useful measures on a standard scale.

Rather different is the problem of choosing between two, or a few, separate families of models. Here the symmetric roles of the hypotheses seem to me to make significance testing very artificial. It would be better to adopt fair empirical comparisons, using cross-validation or bootstrap methods, or a full-fledged Bayesian calculation. The latter requires careful choice of prior distributions within each model to avoid inconsistencies.

Significance tests will sometimes be used for a nuisance factor, preliminary to the main test, as with the initial test for a cross-over effect in a comparative trial with cross-over design. Racine, Grieve, Fluhler, and Smith (1986) recently demonstrated the clear merits of a Bayesian approach in this context. If significance tests are to be useful, then they should have validity independent of the values of identifiable nuisance factors.

In general, for problems where the usual null hypothesis defines a special value for a parameter, surely it would be more informative to give a confidence range for that parameter. Note that some significance tests are not compatible with efficient confidence statements, simply

*David V. Hinkley is Professor, Department of Mathematics, University of Texas, Austin, TX 78712.

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

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.

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