DICKEY, J. M. (1977). Is the tail area useful as an approximate Bayes factor? *J. Amer. Statist. Assoc.* **72** 138–142.

DICKEY, J. M. (1980). Approximate coherence for regression model inference— with a new analysis of Fisher's Broadback Wheatfield example. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) 333–354. North-Holland, Amsterdam.

EDWARDS, W., LINDMAN, H. and SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70** 193–242. Reprinted in *Robustness of Bayesian Analysis* (J. B. Kadane, ed.). North-Holland, Amsterdam, 1984.

GÓMEZ VILLEGAS, M. A. and DE LA HORRA NAVARRO, J. (1984). Aproximacion de factores Bayes. *Cuad. Bioestadist.* **2** 355–361.

GOOD, I. J. (1950). *Probability and the Weighing of Evidence.* Charles Griffin, London.

GOOD, I. J. (1958). Significance tests in parallel and in series. *J. Amer. Statist. Assoc.* **53** 799–813.

GOOD, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press, Cambridge, Mass.

GOOD, I. J. (1967). A Bayesian significance test for the multinomial distribution. *J. Roy. Statist. Soc. Ser. B* **29** 399–431.

GOOD, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* Univ. Minnesota Press, Minneapolis.

GOOD, I. J. (1984). Notes C140, C144, C199, C200 and C201. *J. Statist. Comput. Simulation* **19**.

GOOD, I. J. (1985). Weight of evidence: A brief survey. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 249–270. North-Holland, Amsterdam.

GOOD, I. J. (1986). A flexible Bayesian model for comparing two treatments. *J. Statist. Comput. Simulation* **26** 301–305.

HILDRETH, C. (1963). Bayesian statisticians and remote clients. *Econometrica* **31** 422–438.

HILL, B. (1982). Comment on "Lindley's paradox," by G. Shafer. *J. Amer. Statist. Assoc.* **77** 344–347.

HODGES, J. L., JR. and LEHMANN, E. L. (1954). Testing the approximate validity of statistical hypotheses. *J. Roy. Statist. Soc. Ser. B* **16** 261–268.

JEFFREYS, H. (1957). *Scientific Inference.* Cambridge Univ. Press, Cambridge.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press, London.

JEFFREYS, H. (1980). Some general points in probability theory. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) 451–453. North-Holland, Amsterdam.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley, New York.

LEMPERS, F. B. (1971). *Posterior Probabilities of Alternative Linear Models.* Univ. Rotterdam Press, Rotterdam.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192.

LINDLEY, D. V. (1961). The use of prior probability distributions in statistical inference and decisions. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 453–468. Univ. California Press.

LINDLEY, D. V. (1965). *An Introduction to Probability and Statistics from a Bayesian Viewpoint* **1, 2**. Cambridge Univ. Press, Cambridge.

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213.

PRATT, J. W. (1965). Bayesian interpretation of standard inference statements (with discussion). *J. Roy. Statist. Soc. Ser. B* **27** 169–203.

RAIFFA, H. and SCHLAIFFER, R. (1961). *Applied Statistical Decision Theory.* Division of Research, Graduate School of Business Administration, Harvard Univ.

RUBIN, H. (1971). A decision-theoretic approach to the problem of testing a null hypothesis. In *Statistical Decision Theory and Related Topics* (S. S. Gupta and J. Yackel, eds.). Academic, New York.

SHAFER, G. (1982). Lindley's paradox (with discussion). *J. Amer. Statist. Assoc.* **77** 325–351.

SMITH, A. F. M. and SPIEGELHALTER, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser. B* **42** 213–220.

SMITH, C. A. B. (1965). Personal probability and statistical analysis. *J. Roy. Statist. Soc. Ser. A* **128** 469–499.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Economics.* Wiley, New York.

ZELLNER, A. (1984). Posterior odds ratios for regression hypotheses: General considerations and some specific results. In *Basic Issues in Econometrics* (A. Zellner, ed.) 275–305. Univ. Chicago Press, Chicago.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. University Press, Valencia.

Comment

D. R. Cox

The use and misuse of significance tests has been a popular topic of comment for many years and from many points of view. Although the matter could hardly be said to be underdiscussed in the recent literature, Professors Berger and Delampady have made a valuable addition to that literature by their careful account of the relation between P-values and the posterior probabilities of "precise" null hypotheses, a matter first raised many years ago by H. Jeffreys. The extension of the discussion to include broad classes of priors is particularly striking.

For those taking an eclectic view of statistical theory the comparison of different approaches to the same or similar problems is important, sometimes soothing and occasionally constructively alarming. What is one to make of the present comparisons? The authors are in no doubt.

(i) "Rejoinder 5. P-values have a valid frequentist interpretation. This rejoinder is simply not true" (Section 4.5).

D. R. Cox is Professor of Statistics, SERC Senior Research Fellow, Department of Mathematics, Imperial College, London SW7 2BZ, United Kingdom.

(ii) "... inclusion of all data "more extreme" than the actual x_0 is a curious step, and one which we have seen no remotely convincing justification for" (Section 4.6).

(iii) "... formal use of P-values should be abandoned. Almost anything will give a better indication ... " (Section 5).

Extracts out of context are potentially misleading but I hope that the above give a fair basis for discussing the implications of the paper. Note also the absence of the following.

(iv) Rejoinder 8 or 4′. Attempts to force formal problems of statistical inference into an exclusively Bayesian mold may give misleading answers.

It seems to me that the authors' discussion shows that sometimes a substantial improvement on the use of P-values will be possible but that nevertheless (i) is based on a conceptual error on their part, that this leads them to their unwise view (ii), that (iii) is a considerable exaggeration and that they have underestimated the force of (iv).

First it is important to stress, as do the authors, that there are several kinds of null hypothesis and that many encountered in applications are dividing hypotheses, the question of interest being whether the direction of such and such an effect is reasonably firmly established. Here the apparent disagreements between different approaches are normally minor.

Turning now to points of disagreement, I think that in (i) the authors are confused over the difference between a statement that gives a concept its hypothetical operational meaning, and a statement intended as a basis for direct use. In the first sense the hypothetical long run frequency interpretation of a significance level seems totally clear and unambiguous; it can be argued to be useless but not to be meaningless or invalid. Think, for example, of a geophysicist who finds it useful to work with the acceleration due to gravity at sea level under the summit of Mt. Everest. If challenged as to what this means, he might talk about excavating a small chamber at the appropriate point under the mountain, dropping a small ball, ... The fact that he could not carry out such an experiment and probably would not want to even if he could does not make the concept and explanation meaningless. Whether it is a fruitful idea can be discovered only by trying to use it.

A similar comment applies to point (ii) above. A standard defense of tail areas and P-values is that if we were to regard the data under analysis as just decisive against H_0 then we would have to regard more extreme samples also as evidence against H_0. Therefore, P is the probability of declaring there to be evidence against H_0 when it is in fact true and the data under analysis are regarded as just decisive. This is entirely hypothetical, but that does not make it meaningless. The authors are, of course, entitled to say that they regard this as "not remotely convincing" but it is then a little hard to give a reasoned reply. I can only say that it is a quite clear and precise interpretation, that it seems to me sometimes a relevant notion and that its admittedly somewhat contorted character is, in some contexts, a rather small price to pay for operating in a minimal formulation.

Now consider the final points (iii) and (iv). If we have H_0 and a parametric family of alternatives, the authors' calculations show that the posterior probability of H_0 can be quite appreciable even when P is in some conventional sense quite small. In particular if the alternative is simple, and one disregards the possibility that neither hypothesis holds I certainly agree that a likelihood ratio is much more incisive than a tail area. But this is not really typical of situations for which significance tests are most useful. The authors have summarized such positions well in Section 5. We have a reasonably precise idea of a null hypothesis. We consider that the null hypothesis may be inadequate in some way we, at the moment, do not specify in any detail. If we find clear evidence against the null hypothesis we may well explore possibilities in much more detail.

The authors' arguments suggest that as regards the support of some *particular* kind of alternative we may do well to be cautious, but that is not really what the conventional significance test is about; it is to serve as a general warning that something is wrong (or not), not as explicit support for a particular alternative explanation. Thus, such tests have a very limited aim and often one should be doing something more strongly focused, but that does not make the P-value misleading or useless. To turn the authors' argument on its head, how can they be universally happy with checks on the adequacy of hypotheses that are so cautious as to give false signals of inadequacy only extremely rarely, with consequent low sensitivity?

In summary, it seems to me that the paper is a valuable and thought-provoking one, but that the conclusion that P-values have no role at all is wrong.