Delampady, that "formal use of P-values should be abandoned" (Section 5) is based on a faulty premise, the premise that the Bayesian point null calculation with large π_0 is infallible and appropriate in all point null testing problems. Because this is far from the case, the use of P-values should not be abandoned.

ACKNOWLEDGMENTS

This research was supported by Grant DMS-85-01973 from the National Science Foundation and United States Army Research Office Grant DAAG 29-82-K-0168 at Florida State University.

Comment

Joseph B. Kadane

Testing precise hypotheses played a large role in my statistical education at Stanford. When I left Stanford to teach at Yale in 1966, the book I regarded as fundamental to statistical theory, the one I most wanted to teach, was Lehmann's (1959) on hypothesis testing. My view was that learning about the simplest decision case, where there are only two decisions, would be useful to developing a deeper understanding of more complex decision problems.

Two surprises occurred at Yale. The first was that I met Jimmie Savage and started to learn about Bayesian statistics. The second was that when I tried to use my favorite statistical method on data, trouble ensued. In some joint work with a sociologist, Kadane, Lewis and Ramage (1969), we were examining whether a certain theory predicting frequency of participation in group discussions fit the data. The difference was significant at the .05 level, the .01 level and in fact the 10^{-6} level. I had to think about whether I would be more impressed if it were significant at the 10^{-13} level, and had to conclude that I would not. Ultimately, we found a way to plot the theory and the data together and found the theory to be reasonable but not terribly impressive as a summary of the data. The problem, of course, was that we had too much data, so the statistical significance test was uninformative.

A second difficulty occurred later when I was on the staff of the Center for Naval Analyses. A machine had been developed and tested extensively in a laboratory. It was then tested in the field, and the draft of the results said that the machine was not working differently in the field than it was in the laboratory. However, there were only five observations, each costing a million dollars to collect. The machine was working about 75% as well in the field as it did in the laboratory.

In thinking about these two examples, it became clear to me that what drove the significance test is the sample size: with a large data set everything is significant, but with a small data set, nothing is significant. Having less complex measures of sample size, the usefulness of significance testing was in serious doubt.

Of course, in neither case did the null hypothesis have any special claim on my belief. Because I did not believe the null hypothesis anyway, the calculation of the probability that some statistic would be this or more extreme were the null hypothesis true, is not informative to me. Estimating anything reasonable—like the distance of the data from the theory in the group discussion problem or the degree of degradation in the field in the Navy problem—seems much more sensible.

For the last 15 or so years I have been looking for applied cases in which I might have some serious belief in a null hypothesis. In that time I found only one. An astrologer of my acquaintance believed she could predict on the basis of people's birthdates who is likely to have a drug problem. I arranged for the obtaining of birthdates of persons who were in a Veterans Administration drug treatment program, and of persons under the care of a physician and known by him not to have drug problems. The dates were shuffled up and sent to the astrologer. She rated each person on a one to nine scale of the likelihood of having a drug problem. The data were analyzed using the Mann-Whitney statistic as an estimate, and showed that a randomly chosen Veterans Administration patient had a 48.5% probability of being rated more likely to have a drug problem that a randomly chosen drug-free patient. Thus the astrologer was predicting slightly worse than chance. Even in this case I find the estimate, 48.5%, more meaningful than I would a test of a null hypothesis (should it be one-tailed or two-tailed?).

My conclusion now from these experiences is that

Joseph B. Kadane is the Leonard J. Savage Professor of Statistics and Social Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. These comments were written while the author was on sabbatical leave at the Center for Advanced Study in the Behavioral Sciences, Stanford, California.

the technique of testing hypotheses is vastly overrated in statistics as a method. It isn't so much that the classical methods give the wrong answers, as Berger and Delampady correctly show, as it is that I find the problem ill-suited to help me do statistics better. Thus, I find myself in agreement with Berger and Delampady that "when testing precise hypotheses, formal use of P-values should be abandoned." On the other hand, I do not expect to test a precise hypothesis as a serious statistical calculation.

ADDITIONAL REFERENCES

KADANE, J. B., LEWIS, G. and RAMAGE, J. (1969). Horvath's theory of participation in group discussion. *Sociometry* **32** 348–361.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

Rejoinder

James O. Berger and Mohan Delampady

We are grateful to the discussants for their comments. All raise interesting issues that are highly deserving of discussion. As usual, we will focus on disagreements in our rejoinder.

REPLY TO COX

Professor Cox questions our argument that P-values do not have a valid frequentist interpretation, stating that the "hypothetical long-run frequency interpretation of a significance level seems totally clear and unambiguous." Over many years of trying to understand what makes a valid frequentist interpretation, we have come to agree with Neyman's view that one must have a stated accuracy criterion, a stated procedure and determine the expected accuracy of the procedure in repeated use; thus, an $\alpha = .05$ level test will indeed reject true nulls only 5% of the time in repeated use. A P-value has no such *real* frequentist interpretation. It has various pseudofrequentist interpretations (cf. Cox and Hinkley, 1974), but these are somewhat contorted so that their impact, or persuasiveness, is much less than that of the *real* frequentist justification. Also, a thorough study of our Example 6 is, we feel, very important in understanding the role of frequentism here.

The reaction of Cox to our claim, that ". . . inclusion of all data 'more extreme' than x_0 is a curious step and one we have seen no remotely convincing justification for," is to say that he finds the reasoning clear and precise and at least sometimes relevant. He, of course, is well aware of the many examples in statistics (some due to Cox himself) where inclusion of "other data" in the calculation leads to nonsense. We submit that this is one of those situations, and indeed can marshall (following Jeffreys) purely intuitive arguments against including more extreme data: is it really fair to H_0 to hurl against it not just the (mild) evidence x_0, but also all the much stronger "extreme" values, when these extreme values *did not occur*?

We, for the most part, agree with the remaining comments of Cox. Our statement that "formal use of P-values should be abandoned" was directed to the formal use of P-values in providing quantitative measures of doubt of H_0. At the beginning of Section 5 we agreed that the informal use of P-values "as a general warning that something is wrong (or not) . . . " (to use Cox's phrase) is perhaps reasonable; this informal use in data analysis may well justify the teaching and consideration of P-values.

In regard to "sensible uses of P-values," it is worth considering an earlier comment of Cox to the effect that for "dividing hypotheses . . . the apparent disagreements between different approaches are normally minor." We used to think this, but the discussion of Carl Morris to Berger and Sellke (1987) shows that such may well not be so.

Finally, our response to Cox's Rejoinder 8 or 4′ is what would be expected of Bayesians: We feel that using the Bayesian paradigm will give misleading answers less often than use of alternative paradigms.

REPLY TO EATON

We agree with just about everything in Professor Eaton's discussion, leaving us little to do but applaud the further insights provided. The objectivity issue is indeed a fundamental concern. Eaton argues that objectivity is a vague, ill-defined concept, and may not exist. We agree; indeed, one of the major purposes of the paper was to show that Opinion 2 in the introduction is wrong. Testing a precise hypothesis is a situation in which there is *clearly* no objective Bayesian analysis and, by implication, no sensible objective analysis whatsover. In other problems, arguments about whether noninformative priors are, or are not, objective tend to be inconclusive, but here there simply is *no* prior that can even be called noninformative.

Although the precise hypothesis testing scenario was used to demonstrate that objectivity is at least