

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  $\pi_0$  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.

## 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

---

*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.*

## 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.

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 than 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