where d_1 means rejecting H_0 and b is a constant. An appropriate distance δ between θ and θ_0 in this case is the standardized square distance (Mahalanobis distance)

$$\delta(\theta, \theta_0) = \{(\theta - \theta_0)/\sigma\}^2,$$

which happens to be twice the Kullback-Leibler divergence between the $N(\theta, \sigma^2)$ and the $N(\theta_0, \sigma^2)$ distributions. According to the discussion above, we will reject H_0 if and only if

$$E[\delta(\theta, \theta_0) \mid \mathbf{x}] > \delta_0.$$

If we take the usual "objective" prior for this problem, $\pi(\theta) \propto 1$, then the posterior distribution of θ is simply $N(\bar{x}, \sigma^2/n)$ so that

$$\begin{aligned} U(\mathbf{x}) &= E[\delta(\theta, \theta_0) \mid \mathbf{x}] \\ &= (1/n) + (\bar{x} - \theta_0)^2/\sigma^2 = (1 + T^2)/n \end{aligned}$$

where T is given in Example 1. Then we will reject H_0 whenever $T^2 > c(n) = n\delta_0 - 1$.

We could explicitly seek an analogy with the classical methodology and thus select δ_0 to be the $1 - \alpha$ quantile of the sampling distribution of $U = U(\mathbf{X})$ under the null hypothesis, where α is the level of significance (not the P-value as in Example 1). In this case, with this *particular* value of n, we would reproduce the frequentist test procedure. But if the value of n changes, δ_0 still must have the same value, so that $c(n)$ must change. Thus, the frequentist rule of choosing $c(n)$ so that the test has size α can have a Bayesian interpretation as long as α changes accordingly with the results above. Of course, this example is just a particular case of the problem studied in Ferrandiz (1985).

ADDITIONAL REFERENCES

BAYARRI, M. J. (1985). A Bayesian test for goodness-of-fit. Technical Report, Departamento de Estadística e Investigación Operativa, Univ. Valencia.

DEGROOT, M. H. and MEZZICH, J. E. (1985). Psychiatric statistics. In *A Celebration of Stastistics: The ISI Centenary Volume* (A. C. Atkinson and S. E. Fienberg, eds.) 145–165. Springer, New York.

FERRANDIZ, J. R. (1985). Bayesian inference on Mahalanobis distance: An alternative approach to Bayesian model testing. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 645–654. North-Holland, Amsterdam.

ZELLNER, A. (1980). Statistical analysis of hypotheses in economics and econometrics. *Proc. Amer. Statist. Assoc. Bus. Econ. Statist. Sec.* 199–203.

Comment

George Casella and Roger L. Berger

We congratulate Berger and Delampady on an informative paper. However, we do not believe that the point null testing problem they have considered reflects the common usage of point null tests. Their main thesis is that the frequentist P-value overstates the evidence against the null hypothesis although the Bayesian posterior probability of the null hypothesis is a more sensible measure. A second point of their paper is that point null hypotheses are reasonable approximations for some small interval nulls. We disagree with both of these points.

The large posterior probability of H_0 that Berger and Delampady compute is a result of the large prior probability they assign to H_0, a prior probability that is much larger than is reasonable for most problems in which point null tests are used. Replacing a large prior probability for a point by an equally large prior probability for a small interval about the point does not remedy the problem. It only replaces one unrealistic problem with another. We will argue that given a reasonably small prior probability for an interval about the point null, the posterior probability and the P-value do not disagree. Before moving to the main points of our rejoinder, however, we would like to make a general comment.

Contrary to what Berger and Delampady would have us believe, a great many practitioners should not be testing point nulls, but should be setting up confidence intervals. Interval estimation is, in our opinion, superior to point null hypothesis testing, Rejoinder 3 of Berger and Delampady notwithstanding. However, we will not argue about the appropriateness of the test of a point null. Instead, we will argue the following: Given the common problems in which point null tests are used, the Bayesian measure of evidence, as exemplified by equation (4) of Berger and Delampady is not a meaningful measure. In fact, it is not the case that P-values are too small, but rather that Bayes point null posterior probabilities are much too big!

George Casella is Associate Professor, Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, New York 14853. Roger L. Berger is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695.

First we will discuss types of precise null hypotheses and suggest that the type considered by Berger and Delampady is not common. Then we will make some comments regarding interval null hypotheses.

In Section 5, they describe two types of precise hypotheses. They point out that their results only apply to the second type. But they have ignored a third type, the type that describes the most common usage of point null tests. Consider the following three types; (1) and (2) were the two mentioned by Berger and Delampady.

(1) Precise hypotheses that are just stated for convenience and have no special prior believability.

(2) Precise hypotheses that do correspond to a concentration of prior belief.

(3) Precise hypotheses that describe a unique, interesting feature of the population but that have no special prior believability.

We will discuss each of these types.

As an example of type (1), Berger and Delampady seem to suggest a situation in which a one-sided test is appropriate, but a two-sided point null test is used. Another example might be a one-sided problem in which H_0: $\theta = \theta_0$ rather than the appropriate H_0: $\theta \leq \theta_0$ is used. (Casella and Berger (1987) point out that this convenient restatement creates a bias *toward* H_0 in a Bayesian analysis.) In either case the hypotheses have not been properly formulated. Our concern should not be to analyze these misspecified problems, but to educate the user so that the hypotheses are properly formulated. So although, as Berger and Delampady admit, the P-value might be a reasonable measure of evidence in this type of problem, we should be more concerned with ensuring that these *convenient* hypotheses are not tested.

Type (2) hypotheses are the type considered in Berger and Delampady. In fact, in their tables (Tables 1, 4, 5, 6, 7 and 8) in which P-values and $P(H_0 \mid x)$ are compared, $\pi_0 = ½$ is used. Most researchers would not put a 50% prior probability on H_0. The purpose of an experiment is often to disprove H_0 and researchers are not performing experiments that they believe, *a priori*, will fail half the time! We would be surprised if most researchers would place even a 10% prior probability on H_0. We hope that the casual reader of Berger and Delampady realizes that the big discrepancies between P-values and $P(H_0 \mid x)$ that are reported in the tables are due to a large extent to the large value of $\pi_0 = ½$ that was used. Statements of Berger and Delampady, such as "when testing precise hypotheses, formal use of P-values should be abandoned," must be qualified to apply only to type (2) hypotheses with unusually large values of π_0.

We believe that most point null hypotheses that are tested are of type (3). If H_0 were true, then the population would have some unique, interesting feature. But the researcher does not believe, *a priori*, that this feature exists and, in fact, probably expects to show that H_0 is not true. The following two examples, we believe, encompass many point null tests that are done. In neither example is the researcher likely to believe that H_0 is true. In the first example, $\theta = \mu_1 - \mu_2$, the difference between two population means and H_0: $\theta = 0$ is tested with a paired difference or independent samples test. It would be a very interesting situation if μ_1 were to equal μ_2, but the researcher does not typically believe that this is even approximately true, much less exactly true. In the second example, H_0: $\beta_i = 0$ is tested where β_i is a regression coefficient. Again, it would be an important feature of the population if H_0 were true. It would indicate that the independent variable x_i has no effect on the response variable. But the researcher does not place a high prior probability on H_0. Indeed, x_i, probably would not have been included in the experiment if the researcher thought that it was highly likely that x_i, was unrelated to the response variable. We believe that these examples typify the common usages of point null tests and, as Berger and Delampady admit in Section 5, P-values are reasonable measures of evidence when there is no *a priori* concentration of belief about H_0.

Much of their paper concerns testing an interval null, H_0: $\lvert \theta - \theta_0 \rvert \leq \varepsilon$, rather than testing a point null. There are two points regarding interval nulls on which we would like to elaborate. These are: (a) The Bayesian test of a point null, with $\pi_0 = ½$, cannot be approximated by a test of an interval null hypothesis in problems unless there is a high concentration of prior belief about the point null. (b) Bayesian posterior probabilities of interval null hypotheses are *quite close to P-values* when the prior probability of H_0 is reasonably small.

They show that the Bayesian measures of evidence are about the same if one tests H_0: $\lvert \theta - \theta_0 \rvert \leq \varepsilon$ or if one tests H_0: $\theta = \theta_0$ if ε is sufficiently small. In both cases the prior probability assigned to H_0 is π_0. They say that this refutes the claim that the discrepancies between P-values and $P(H_0 \mid x)$ are caused by assignment of mass to a single point. But we do not believe that assignment of a large probability, say $\pi_0 = ½$, to a tiny interval is much more realistic than assignment of π_0 to the point $\theta = \theta_0$. In the above example, not only does the researcher typically not assign probability ½ to the hypothesis $\mu_1 = \mu_2$ but also does not assign probability ½ to the interval $\lvert \mu_1 - \mu_2 \rvert \leq \varepsilon$ where ε is small. The point is the same as above. The hypothesis $\mu_1 = \mu_2$ is of interest not because there is high prior probability concentrated about it but because of the interesting feature of the populations it describes. We see the Berger and Delampady results mainly of interest to the Bayesian hypothesis tester who assigns probability π_0 to H_0: $|\theta - \theta_0| \le \varepsilon$ and who can simplify his calculations by approximating this problem with the problem in which probability π_0 is assigned to the point null $\theta = \theta_0$. To see how small this interval must be for the approximation to be valid, note that if $n = 25$ and $\varepsilon^* = .4$ (a medium value from Table 3 of Berger and Delampady) then ε must be less than $.80\sigma$.

If the Bayesian assigns prior probability π_0 to H_0: $|\theta - \theta_0| \le \varepsilon_0$ then ε_0 should not (indeed, cannot) depend on n, the sample size. We believe the relevant calculation in this case is the one done by Berger and Delampady in Section 2.3, where they show that $P(H_0 \mid \bar{x}_n) \to \alpha$ as $n \to \infty$ where the P-value associated with $\bar{x}_n$ is α. So the Bayesian can use the P-value as an approximate posterior probability for large n, regardless of the value of π_0.

In the typical case in which the prior probability assigned to H_0: $|\theta - \theta_0| \le \varepsilon$ is small, this hypothesis may still be of interest. It says that the population is "close" to having the unique feature associated with $\theta = \theta_0$. But in this case the P-value and $P(H_0 \mid x)$ do not display the wide discrepancies that occur when the prior probability assigned to H_0 is large. Consider the following comparison of P-values and $P(|\theta| \le \varepsilon \mid x)$, which can be thought of as an amendment to Table 2 of Berger and Delampady. Here, ε^* is taken from their Table 2 and the probabilities are calculated according to $X \mid \theta \sim n(\theta, 1)$, $\theta \sim n(0, 2^2)$.

Table 1 shows that the Bayesian interval measure is quite close to the P-value, which supports our point (b). In Table 1, $\varepsilon = \varepsilon^*$ was just chosen as a typical small interval. In fact, for a range of values of ε, and a range of values of x, this phenomenon persists. The P-value and $P(|\theta| \le \varepsilon \mid x)$ are relatively close together, although $P(\theta \doteq 0 \mid x)$ is far from both of them. This is illustrated in Figure 1.

TABLE 1
Comparison of P-values and $P(|\theta| \le \varepsilon \mid x)$

x	1.645	1.96	2.576	2.807	3.29	3.89
P-value	.10	.05	.01	.005	.001	.0001
$\varepsilon = \varepsilon^*$	.257	.221	.173	.160	.138	.117
$P(\|\theta\| \le \varepsilon \mid x)$	.079	.043	.011	.006	.002	.0003

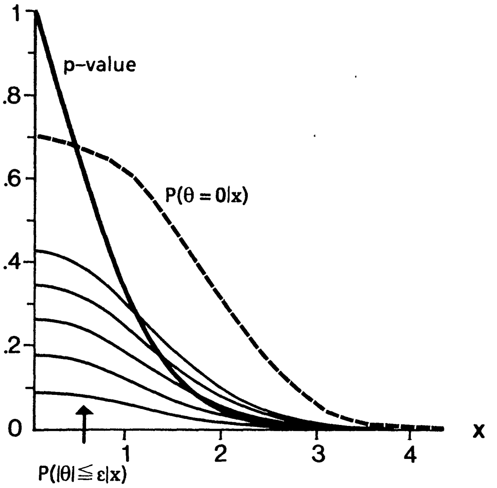

FIG. 1. For $X \mid \theta \sim n(\theta, 1)$, P-value is the two-sided P-value. $P(\theta = 0 \mid x)$ is calculated using a point mass of ½ at $\theta = 0$, and $n(0, 2^2)$ prior elsewhere. $P(|\theta| \le \varepsilon \mid x)$ uses only the $n(0, 2^2)$ prior and is shown for $\varepsilon = .1, .2, .3, .4, .5$. The curves are increasing in ε.

TABLE 2
Interval posterior probabilities for the binomial

α	n	x	θ_0	P_c	$P(\|\theta - \theta_0\| \le \varepsilon \mid x)$
.0090	50	11	.40	.0981	.030
.0100	20	9	.20	.1771	.053
.0101	20	14	.40	.1064	.025
.0118	20	4	.50	.0858	.021
.0120	45	10	.10	.2211	.145
.0493	50	16	.20	.3313	.170
.0505	15	1	.30	.1956	.055
.0507	25	3	.30	.2414	.069
.0541	40	10	.40	.3016	.102
.0556	15	4	.10	.4223	.214
.0960	15	6	.20	.4123	.159
.0980	25	5	.10	.4779	.341
.0987	30	20	.50	.3565	.117
.1000	35	15	.30	.4328	.200
.1011	10	7	.40	.3458	.095
.1094	10	2	.50	.3163	.084

The combination of our belief that the testing of a point null or a small interval null does not usually imply a high prior probability concentrated at H_0 and our numerical calculations to support our point (b) lead us to conclude that the fault is not with the P-value, but with the Bayesian point-mass calculation. The agreement between P-values and interval null probabilities is not restricted to the normal case, but also occurs in the binomial case. Consider Table 2, an amendment to Table 7 of Berger and Delampady. In Table 2, $X \mid \theta \sim$ binomial (n, θ), and the first five columns are the same as Table 7 of Berger and Delampady. The interval posterior probability is calculated using a beta $[c\theta_0, c(1 - \theta_0)]$ prior, with $c = 5$. The value of ε was .05.

In summary, we have, at the very least, demonstrated that there exist legitimate criticisms of the Bayesian point null calculations, and dismissing P-values based on a lack of agreement with the point null calculations is unjustified. Moreover, there is agreement between P-values and Bayesian interval null calculations in the more typical situation in which small prior probability is assigned to H_0. So the very argument that Berger and Delampady use to dismiss P-values can be turned around to argue *for* P-values. The recommendation of Berger and

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.