Discussion of

"TESTING PRECISE HYPOTHESES"

By

James Berger and Mohan Delampady

BU-931-M

May, 1987

George Casella[1]

Cornell University

Roger L. Berger

North Carolina State University

---

[1]Research supported by NSF Grant No. DMS 850-1973.

We congratulate Berger and Delampady (B&D) 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 while 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 B & D 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. And 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 B&D 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, B&D Rejoinder 3 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 B&D's equation (4) 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!

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

In Section 5, B&D 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 B&D.

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), B&D 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 & 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 B&D 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 B&D. In fact, in their tables (Tables 1, 4, 5, 6, 7, and 8) in which P-values and $P(H_0|x)$ are compared, $\pi_0=\frac{1}{2}$ 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 B&D realizes that the big discrepancies between P-values and $P(H_0|x)$ that are reported in the tables are due to a large extent to

the large value of $\pi_0 = \frac{1}{2}$ that was used. Statements of B&D, 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 B&D admit in Section 5, P-values are reasonable measures of evidence when there is no apriori concentration of belief about H_0.

Much of B&D's paper concerns testing an interval null, $H_0: |\theta - \theta_0| \leq \epsilon$, 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 $\pi_0 = \frac{1}{2}$, 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.

B&D show that the Bayesian measures of evidence are about the same if one tests $H_0:|\theta-\theta_0| \leq \epsilon$ 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 $P(H_0|x)$ are caused by assignment of mass to a single point. But we do not believe that assignment of a large probability, say $\pi_0=\frac{1}{2}$, 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 $\frac{1}{2}$ to the hypothesis $\mu_1=\mu_2$ but also does not assign probability $\frac{1}{2}$ to the interval $|\mu_1-\mu_2| \leq \epsilon$ 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 B&D results mainly of interest to the Bayesian hypothesis tester who assigns probability π_0 to $H_0:|\theta-\theta_0| \leq \epsilon$ 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 $\epsilon^*=.4$ (a medium value from Table 3 of B&D) then ϵ must be less than $.08\sigma$.

If the Bayesian assigns prior probability π_0 to $H_0:|\theta-\theta_0| \leq \epsilon_0$ 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 B&D in Section 2.3 where they show that $P(H_0|\bar{x}_n) \rightarrow \alpha$ as $n \rightarrow \infty$ 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 $H_0:|\theta-\theta_0| \leq \epsilon$ 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 $P(H_0|x)$ 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 $P(|\theta| \leq \epsilon|x)$, which can be thought of as an amendment to Table 2 of B&D. Here, ϵ^* is taken from Table 2 of B&D, and the probabilities are calculated according to $X|\theta \sim n(\theta,1)$, $\theta \sim n(0,2^2)$.

Table 1: Comparison of P-values and $P(|\theta| \leq \epsilon|x)$

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

Table 1 shows that the Bayesian interval measure is quite close to the P-value, which supports our point (b). In Table 1, $\epsilon=\epsilon^*$ 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 $P(|\theta| \leq \epsilon|x)$ are relatively close together, while $P(\theta=0|x)$ is far from both of them. This is illustrated in the following Figure 1.

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

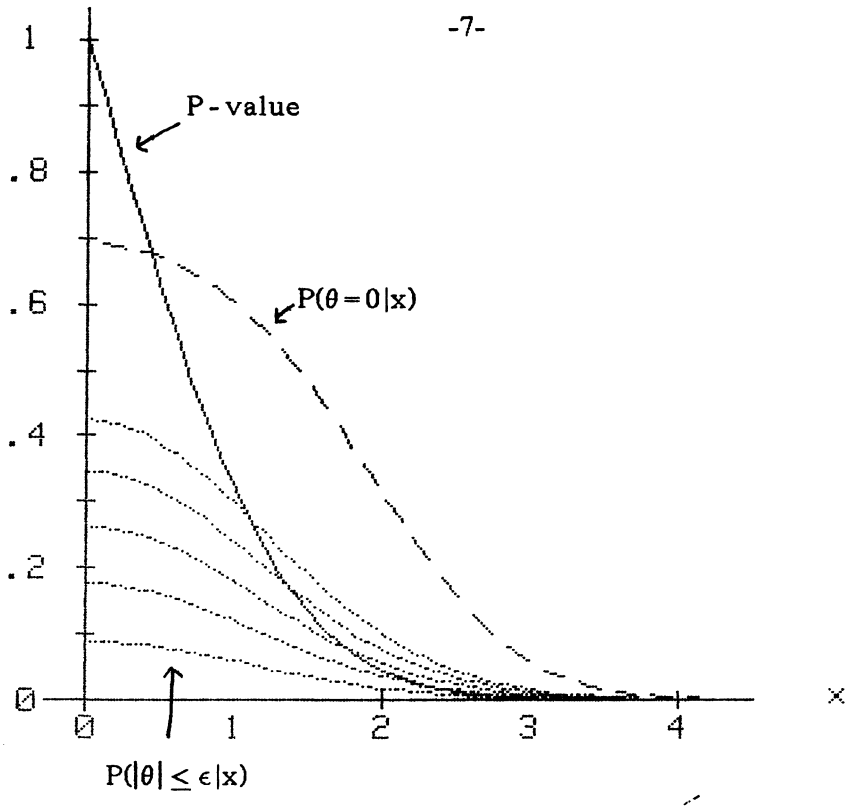


Figure 1: For $X|\theta \sim n(\theta,1)$, P-value is the two-sided P-value. $P(\theta=0|x)$ is calculated using a point mass of $\frac{1}{2}$ at $\theta=0$, and $n(0,2^2)$ prior elsewhere. $P(|\theta|\leq\epsilon|x)$ uses only the $n(0,2^2)$ prior, and is shown for $\epsilon=.1,.2,.3,.4,.5$. The curves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

amendment to Table 7 of B&D. In Table 2, $X|\theta \sim$ binomial (n,θ), and the first five columns are the same as B&D's Table 7. The interval posterior probability is calculated using a Beta $\left(c\theta_0, c(1-\theta_0)\right)$ 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 B&D use to dismiss P-values can be

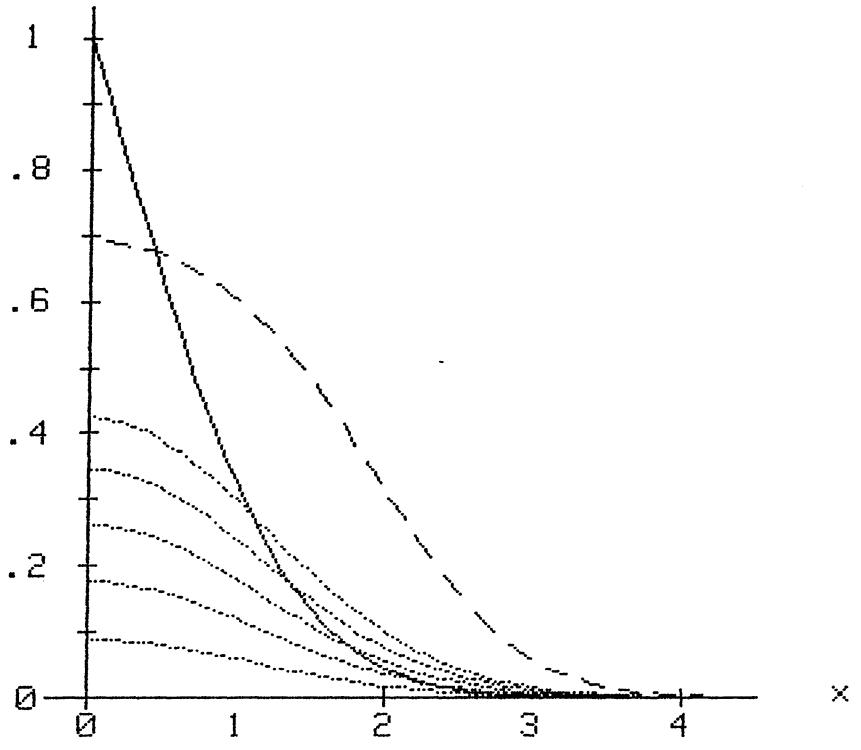
Table 2: Interval Posterior Probabilities for the Binomial

α	n	x	θ_0	$\underline{P}_C$	$P(\lvert\theta-\theta_0\rvert \leq \epsilon \lvert x)$
.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

turned around to argue ***for*** P-values. The recommendation of B&D, 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. Since this is far from the case, the use of P-values should not be abandoned.

Reference

Casella, George and Berger, Roger L. (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem", *Journal of the American Statistical Association*, 82, 106-111.


1
.8
.6
.4
.2
0
0
1
2
3
4
x