Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence: Comment

Author(s): Stephen B. Vardeman
Source: *Journal of the American Statistical Association*, Vol. 82, No. 397 (Mar., 1987), pp. 130-131
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2289136
Accessed: 28-03-2015 14:37 UTC

What should our attitude now be concerning P values? Berger and Sellke note that nonstatisticians tend to confuse the P value and the posterior probability of the null hypothesis. As pointed out in Good (1984), even the most respected statisticians can make the same mistake. The present works reinforce the distinction between sampling probability and posterior probability.

It has long seemed to me that the P value reports an interesting fact about the data. I once speculated to Dennis Lindley that the P value might offer a quicker and cruder form of inference than the Bayes factor. He replied by asking whether what I meant was analogous to comparing an orchestra with a tom-tom.

ADDITIONAL REFERENCES

Dickey, James M. (1976), "Approximate Posterior Distributions," *Journal of the American Statistical Association,* 71, 680–689.
Good, I. J. (1984), "An Error by Neyman Noticed by Dickey" (C209), in "Comments, Conjectures, and Conclusions," *Journal of Statistical Computation and Simulation,* 20, 159–160.

Comment

STEPHEN B. VARDEMAN*

Berger, Sellke, Casella, and Berger deserve our thanks for a most readable and thorough accounting of the problem of comparing p values and posterior probabilities of H_0. They have laid out in very clear fashion the history of the problem, a full array of technical points, and their arguments from the technical points to general conclusions. Their articles should help all of us, card-carrying Bayesians, militant frequentists, and fence-sitters like myself, to sort this issue out to our own satisfaction.

My view from the fence is that in spite of the fact that the articles are well done, there is nothing here very surprising or that carries deep philosophical implications. We all know that Bayesian and frequentist conclusions sometimes agree and sometimes do not, depending on the specifics of a problem. These articles seem to me to reinforce this truism. For example, I read the Casella/Berger Theorem 3.4, the argument behind it, and their subsequent discussion as confirmation that essentially anything can be possible for a posterior probability for H_0, *depending on how one is allowed to move prior mass around on H_0 and H_1*. (Of course, the simplest demonstration that nearly anything can be possible can be made by using arbitrary two-point priors in a composite versus composite case.)

Whether or not a Bayesian analysis can produce a small posterior probability for H_0 is largely a function of whether or not (staying within whatever rules are imposed by the problem structure and restrictions adopted for the prior) one can move the prior mass on H_0 "away from the data," at least as compared with the location of the prior mass on H_1. If this can be done, the posterior probability of H_0 can be made small, otherwise it cannot.

Take, for example, the Jeffreys–Lindley "paradox" discussed by Berger and Sellke. To maintain a p value that is constant with n (i.e., a constant value of t), one must send $\overline{X}_n$ (the data) to θ_0. The nonzero mass on H_0 is trapped at θ_0, while the mass on H_1 is all passed by as $\overline{X}_n \to \theta_0$. Why should anyone then be surprised that the posterior probability assigned to H_0 tends to 1?

Moving to a different point, I must say that I find the "spike at θ_0" feature of the priors used by Berger and Sellke and many before them to be completely unappealing. In fact, contrary to the exposition of Berger and Sellke, I think that the appeal of such priors *decreases* with increasing π_0. Unlike that of Casella and Berger, my objection has nothing to do with "impartiality" (indeed I question whether such a concept can have any real meaning), but is of a more elementary nature. The issue is simply that I do not believe that any scientist, when asked to sketch a distribution describing his belief about a physical constant like the speed of light, would produce anything like the priors used by Berger and Sellke. A unimodal distribution symmetric about the current best value? Probably. But with a spike or "extra" mass concentrated at θ_0? No.

Competent scientists do not believe their own models or theories, but rather treat them as convenient fictions. A small (or even 0) prior probability that the current theory is true is not just a device to make posterior probabilities as small as p values, it is the way good scientists think! The issue to a scientist is not whether a model is true, but rather whether there is another whose predictive power is enough better to justify movement from today's fiction to a new one. Scientific reluctance to change theories is appropriately quantified in terms of a cost structure, not by concentrating prior mass on H_0. In this regard, note that although the "spike at θ_0" priors are necessary to produce nontrivial Bayes rules (i.e., ones that sometimes "accept") for a zero–one type loss structure in the two-sided problem, other competing cost structures do not require them for a Bayesian formulation of the testing

* Stephen B. Vardeman is Professor, Statistics Department and Industrial Engineering Department, Iowa State University, Ames, IA 50011.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

problem to be nontrivial. Consider, for example, a cost structure like

$$\text{cost}(\text{"reject,"}\ \theta) = k_1 - k_2(\theta - \theta_0)^2,$$

$$\text{cost}(\text{"accept,"}\ \theta) = k_3(\theta - \theta_0)^2$$

for positive constants k_1, k_2, and k_3. Here it is clearly possible to have $\Pr[H_0 \text{ is true} \mid \text{data}] = 0$ and at the same time have "accept" be the preferred decision.

A largely nontechnical observation that I feel obliged to make regarding both articles concerns word choice. I would prefer to see loaded words like "biased," "objective," and "impartial" left out of discussions of the present kind, albeit they are given local technical definitions. Too much of what *all* statisticians do, or at least talk about doing, is blatantly subjective for any of us to kid ourselves or the users of our technology into believing that we have operated "impartially" in any true sense. How does one "objectively" decide on a subject of investigation, what variable to measure, what instrument to use to measure it, what scale on which to express the result, what family of distributions to use to describe the response, etcetera, etcetera, etcetera? We can do what seems to us most appropriate, but we can *not* be objective and would do well to avoid language that hints to the contrary.

Having complimented the authors' thoroughness and clarity and expressed some skepticism regarding the depth of the implications that ought to be drawn from their results, I will close these remarks by pointing out what I found to be the most interesting issue they have raised. That is the role of conditioning in the stating of the strength of one's evidence against H_0. I have never been particularly comfortable while trying to convince elementary statistics students that having observed $t = 1.4$ they should immediately switch attention to the event $[|t| \geq 1.4]$. Although I am unmoved to abandon the practice, I do find it interesting that Berger and Sellke see this as the main point at which standard practice goes astray.

Comment

C. N. MORRIS*

These two articles address an extremely important point, one that needs to be understood by all statistical practitioners. I doubt that it is. Let us dwell on a simple realistic example here to see that the Berger–Sellke result is correct in spirit, although case-specific adjustments can be used in place of their lower bounds, and that the Casella–Berger infimum, although computed correctly, is too optimistic for most practical situations.

Example. Mr. Allen, the candidate for political Party A will run against Mr. Baker of Party B for office. Past races between these parties for this office were always close, and it seems that this one will be no exception—Party A candidates always have gotten between 40% and 60% of the vote and have won about half of the elections.

Allen needs to know, for $\theta \equiv$ the proportion of voters favoring him today, whether $H_0 : \theta < .5$ or $H_1 : \theta > .5$ is true. A random sample of n voters is taken, with Y voters favoring Allen. The population is large and it is justifiable to assume that $Y \sim \text{Bin}(n, \theta)$, the binomial distribution. The estimate $\hat{\theta} = Y/n$ will be used.

Question. Which of three outcomes, all having the same p value, would be most encouraging to candidate Allen?

(a) $Y = 15, n = 20, \hat{\theta} = .75$;

(b) $Y = 115, n = 200, \hat{\theta} = .575$;

or

(c) $Y = 1{,}046, n = 2{,}000, \hat{\theta} = .523$.

Facts. The p values are all about .021, with values of $t \equiv (\hat{\theta} - .5)\sqrt{n}/\sigma$, $\sigma \doteq .5$, being 2.03, 2.05, and 2.03. Standard 95% confidence intervals are (.560, .940), (.506, .644), and (.501, .545), respectively. (For the application with $n = 20$, exact binomial calculations are made, and continuity corrections are used for t throughout.)

This problem is modeled as $\hat{\theta} \sim N(\theta, \sigma^2/n)$, given θ, with $\sigma^2 = .25$ known, from binomial considerations. The two hypotheses are taken to be, with $\theta_0 \equiv .5$, $H_0 : \theta < \theta_0$ versus $H_1 : \theta > \theta_0$ (θ_0 is given essentially zero probability). We use the conjugate normal prior distribution, and because of information about past elections, we take $\theta \sim N(\theta_0, \tau^2)$ with $\tau = .05$ so that $\Pr(H_0) = \Pr(H_1) = \frac{1}{2}$ a priori (as both articles assume), and so very probably, .4

* C. N. Morris is Professor, Department of Mathematics and Center for Statistical Sciences, University of Texas, Austin, TX 78712. Support for this work was provided by National Science Foundation Grant DMS-8407876.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods