Taylor & Francis
Taylor & Francis Group

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

Author(s): George Casella and Roger L. Berger

Source: *Journal of the American Statistical Association*, Vol. 82, No. 397 (Mar., 1987), pp. 133-135
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2289138
Accessed: 28-03-2015 14:41 UTC

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Taylor & Francis, Ltd. and American Statistical Association are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

This content downloaded from 200.130.19.157 on Sat, 28 Mar 2015 14:41:54 UTC
All use subject to JSTOR Terms and Conditions

when on "fishing expeditions" with data from unplanned studies, adjustments to t values like those suggested by Berger and Sellke or in formula (3) are mandatory. These facts need to be better understood by the wide population of individuals doing data analyses or interpreting the reports of such analyses. They need to be taught in introductory courses, perhaps when the power of tests is introduced, and should be recognized by the editors of journals that report empirical work in terms of significance tests and p values.

Rejoinder

GEORGE CASELLA and ROGER L. BERGER

We thank Professors Dickey, Good, Hinkley, Morris, Pratt, and Vardeman for their thoughtful and insightful comments. We also thank Professors Berger and Sellke for kindling our interest in this problem.

Before responding to specific points raised by the discussants, we would first like to make some general comments that will, perhaps, make our own beliefs clearer. To some extent we agree with a frequentist colleague of ours who said, upon seeing the title of our article, "Why worry about reconciliation? There is nothing frequentist about a p value." We essentially agree that there is nothing frequentist about a p value, but are concerned, as are Berger and Sellke, that there are a great many statistically naive users who are interpreting p values as probabilities of Type I error or probabilities that H_0 is true. The thesis of Berger and Sellke (B&S) is that these users are grossly wrong in the two-sided case. For us, however, the two-sided case carries along with it many built-in problems, and we considered what seemed to be a more straightforward problem to see if there really were gross deficiencies with p values.

The two-sided case suffers from a certain lack of symmetry that necessitates treating the two hypotheses differently. In particular, the present B&S methodology fixes mass on the null and varies it on the alternative. This is dictated somewhat by the different geometry of H_0 and H_1, but the end result is that there is no way to treat the hypotheses equitably. Therefore, even priors that strive to treat H_0 and H_1 in the same way must contain some subjective input. Of course, even the frequentist model, and hence the p value, may be based on subjective input, but it is only sporting to look for a Bayesian setup that is as impartial (sorry, Professor Vardeman) as possible. The one-sided case presents us with such a setup.

We agree with Professor Good that p values and Bayes factors (or posterior probabilities of H_0) are here to stay. This is one reason why we undertook this study of the relationship between $p(x)$ and $\inf \Pr(H_0 \mid x)$: We wanted to see whether the phenomenon described by B&S in the two-sided problem, namely that the $\inf \Pr(H_0 \mid x)$ is much greater than $p(x)$, also occurs in the one-sided problem. We tried to define precisely conditions under which we could show that the B&S concept of irreconcilability did not hold. Under fairly general conditions in the location parameter model (see Theorem 3.4) we could show that $\inf \Pr(H_0 \mid x) \le p(x)$, and, therefore, the phenomenon of irreconcilability, in general, does not occur in the one-sided testing problem. This leads us to believe that the aforementioned problems with the two-sided setup may be the cause for the discrepancy between the p value and $\Pr(H_0 \mid x)$.

1. REPLY TO DICKEY

We find Professor Dickey accusing us of supporting the thesis of B&S, citing Theorems 3.1 and 3.2 [which show that $p(x) \le \Pr(H_0 \mid x)$ for all priors in the cases considered]. Our main point, however, is that the p value *is* on the boundary of the posterior probabilities, showing that the B&S phenomenon does not necessarily occur in the one-sided case. To support further our thesis of reconcilability, we go on to show that $\inf \Pr(H_0 \mid x) < p(x)$ in many cases, so there is a proper prior for which evidence is reconciled.

It is unclear whether Lindley's comment dissuaded Dickey from his interest in p values, but we feel that there is merit in the concept of the p value as a quick albeit crude form of inference. This is in the spirit of our closing comment that "interpretations of one school of thought can have meaning within the other" (p. 111).

2. REPLY TO GOOD

Professor Good suggests certain interesting parametric classes of priors for the normal mean problem, doing calculations mainly in terms of Bayes factors instead of posterior probabilities. He shows that, for a special case of his priors [$\lambda_0 = \lambda_1 = 0$, $a_0 = a_1 = \tau$, $\Pr(H_0) = \Pr(H_1) = \frac{1}{2}$], reconciliation is possible for τ/σ_n large. But this special case just defines an $n(0, \tau^2)$ prior, so Good's computation with τ/σ_n large is a special case of our computation with $\sigma \to \infty$ in Theorem 3.3. Good, however, does not see this as reconciliation, differentiating between the evidence against $H_0 : \theta \le 0$ and $H_2 : \theta = 0$. This distinction is tangential to the main point, since the p value is always taken as the maximum of $\Pr(X > x \mid \theta)$, the maximum

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

This content downloaded from 200.130.19.157 on Sat, 28 Mar 2015 14:41:54 UTC
All use subject to JSTOR Terms and Conditions

being taken over all θ in H_0. Therefore, the p value is the same for both H_0 and H_2, so although H_0 is not H_2, we have not exaggerated to obtain reconciliation.

3. REPLY TO HINKLEY

The comments of Professor Hinkley offer a number of general ideas about the testing problem, only some of which we agree with. First, we agree that the p value is unambiguously objective, but we do not consider it an error rate. It is precisely for this reason that the p value has come under so much attack from Bayesians [as Jim Berger is quick to point out, $E(p(X) \mid H_0$ is rejected$) = \alpha/2$]. A p value, at best, is a summary of the evidence against H_0 given the data. We agree that it is hopeless to calibrate p values to posterior probabilities, but we were not calibrating. We view $p(x)$ and $\Pr(H_0 \mid x)$ as two interesting and seemingly related measures of statistical evidence. Since they are based on different sets of assumptions, however, a general attempt at calibration is doomed to fail.

We agree with Hinkley's comment that p values provide one convenient way to put useful measures on a standard scale and that the operational interpretation should be relative to the information contained in the data. This concern is also expressed by Good, who proposes standardizing p values to sample sizes of 100. Although we agree that sample size is important in the interpretation of p values, we presently do not endorse these or other attempts at calibration. In fact, we find ourselves very much in agreement with Hinkley's statement concerning confidence ranges and would probably go much further. In a large majority of problems (especially location problems) hypothesis testing is inappropriate: Set up the confidence interval and be done with it!

4. REPLY TO MORRIS

The concerns expressed by Professor Morris share similarities to those of Hinkley and Good, and his simple example proves to be very helpful not only in understanding the relationship between $p(x)$ and $\Pr(H_0 \mid x)$ but also in understanding the essential differences between the one-sided and two-sided problems. The fact that Morris's Equations (1) and (2) describe behavior opposite from that of B&S's Equation (1.1) is very illuminating and shows the large effect that a prior point mass can have.

The election example points out the need for reporting the sample size along with the p value. A good frequentist would always report the probabilities of both Type I and Type II error, and Morris shows us that reporting the sample size along with the p value is somewhat equivalent to this; we thoroughly agree with him. His example also illustrates another of our major concerns about the overuse of hypothesis testing: Setting up the 95% confidence intervals provides an unambiguous choice between (a), (b), and (c).

Morris's calculations further illustrate that the ratio of σ/τ is an important factor in determining whether reconciliation obtains. Our results formalize the way in which reconciliation obtains as the prior information becomes vague with respect to the sample evidence. If the prior information is sharp, the Bayesian and frequentist measures will certainly disagree. This does not make our result irrelevant, however, since we do not say that these measures should agree in all circumstances. Furthermore, in situations with sharp prior information, we would want the measures to disagree, with the relevant measure being chosen according to one's statistical preference.

5. REPLY TO VARDEMAN

The comments of Professor Vardeman perhaps most closely reflect our own views, and part of our article was an attempt to quantify Vardeman's comment that "anything is possible." We too find the "spike at θ_0" distressing and are perhaps more comfortable with a cost structure.

The p value switch from $t = 1.4$ to $t \geq 1.4$ has also been a source of concern for us, because there is no firm frequentist reasoning on which it is based. It no doubt is mimicking the calculation for an α level, but does not have the same theoretical basis that the α-level calculation has. Furthermore, this tail calculation gives obvious bias against H_0 and, for that reason, is not interpretable as an error rate. With appropriate attention to sample size, however, the p value is still valid as a measure of evidence against H_0.

6. REPLY TO PRATT

Saving the best for last, we now turn to Professor Pratt, or in the words of the Beatles, "Mean Mr. Mustard." Pratt believes that the results in our article, besides being rather specialized and not very useful, have already been done by him. Obviously we disagree.

Our main point was that in the one-sided problem the p value does not necessarily overstate the evidence against H_0 in the sense that the p value lies within or on the boundary of a range of reasonable posterior probabilities. Thus an inequality like $\inf \Pr(H_0 \mid x) \leq p(x)$ is not "useless" but, in fact, proves our point.

The simple location model, although admittedly being specialized, is useful for at least two reasons. First, consideration of a simple model can help us gain some understanding about the behavior of these evidence measures; the simple model keeps technical difficulties from masking behavior. Second, the location model, even the normal model with known variance, can provide good approximations to more complicated cases. Many others have considered the location model to be deserving of attention; in particular, Pratt (1965, pp. 182–183) considers this model.

It is not at all clear what was obvious to Pratt in 1965, and perhaps more was obvious to him than to any reader of his paper. In the location model, Pratt stated, "if the prior distribution of θ becomes 'diffuse', then $T - \theta$ and T become independent also, and the p-value becomes exactly the conditional probability that $\theta \leq 0$ given T" (pp. 182–183). No further explanation or proof of this statement is given, so let us look at it more closely and see

This content downloaded from 200.130.19.157 on Sat, 28 Mar 2015 14:41:54 UTC
All use subject to JSTOR Terms and Conditions

some "obvious" implications. First, as Hinkley points out, the p value is completely objective and does not depend on the prior. So as the prior becomes diffuse the p value does not change at all! Perhaps Pratt meant that as the prior becomes diffuse, the posterior probability approaches the p value. But then what is meant by the phrase "becomes diffuse"? In Theorem 3.4, $\sigma \to \infty$ corresponds to the prior becoming diffuse, and we see that $\Pr(H_0 \mid x)$ can converge to any number between 0 and 1 depending on the values of $g(0^-)$ and $g(0^+)$. Therefore, no convergence of $\Pr(H_0 \mid x)$ to $p(x)$ need take place.

In his comment, Pratt qualifies his 1965 statement by eliminating "jagged" priors from considerations. If we interpret jagged to mean discontinuous, then Theorem 3.4 not only points out that only a discontinuity at zero matters but also quantifies the effect of such a discontinuity. In short, Theorem 3.4 gives precise and simple conditions under which the convergence of $\Pr(H_0 \mid x)$ to $p(x)$ will occur.

We believe that there is more value in precise, stylized but verifiable statements than in broad but vague statements that are open to many interpretations, some of which are wrong. This is not to say that intuition is bad, but only that intuition should be backed up by precise theorems. The work of Pratt (1965) is important, with many far-reaching implications—the fact that we are still discussing it 20 years after publication is proof of that. Our work, however, is not contained in Pratt (1965), but rather is, at the least, an extension and formalization of some ideas contained therein.

7. SUMMARY

Bayesians and frequentists may never agree on the appropriate way to analyze data and interpret results, but there is no reason why they cannot learn from one another. Whether or not measures of evidence can be reconciled is probably a minor consideration; understanding what affects a measure of evidence is a major consideration. Some key factors were identified in these articles, more in the comments. Our goal in writing our article was to understand better the similarities and differences between p values and posterior probabilities. With the help of B&S and the discussants we feel that we have succeeded. We hope that the reader has too.

Rejoinder

JAMES O. BERGER and THOMAS SELLKE

We thank all discussants for their interesting comments. Our rejoinder will rather naturally emphasize any disagreements or controversy, and thus will be mainly addressed to the non-Bayesians. We are appreciative of the expressed disagreements, including those of Casella and Berger, since one of our hopes was to provoke discussion of these issues in the profession. These are not dead issues, in the sense of being well known and thoroughly aired long ago; although the issues are not new, we have found the vast majority of statisticians to be largely unaware of them. We should also mention that the commentaries contain many important additional insights with which we agree but will not have the space to discuss adequately. Before replying to the official discussants, we have several comments on the Casella–Berger article.

1. COMMENTS ON THE CASELLA–BERGER ARTICLE

First, we would like to congratulate Casella and Berger on an interesting piece of work; particularly noteworthy was the establishment of the P value as the attained *lower bound* on the posterior probability of the null for many standard one-sided testing situations. It was previously well known that the P value was the limit of the posterior probabilities for increasingly vague priors, but that it is typically the lower bound was not appreciated. And the less common examples, where the lower bound is even smaller than the P value, are certainly of theoretical interest.

Our basic view of the Casella–Berger article, however, is that it pounds another nail into the coffin of P values. To clarify why, consider what it is that makes a statistical concept valuable; of primary importance is that the concept must convey a well-understood and sensible message for the vast majority of problems to which it is applied. Statistical models are valuable, because they can be widely used and yield similar interpretations each time they apply. The notion of 95% "confidence" sets (we here use "confidence" in a nondenominational sense) is valuable, because, for most problems, people know how to interpret them (conditional counterexamples aside). But what can be said about P values? Well, they can certainly be defined for the vast majority of testing problems, but do they give a "sensible message"? In our article we argued that they do not give a sensible message for testing a precise null hypothesis, but one could make the counterargument that this is merely a calibration problem. The P value is after all (usually) a one-to-one monotonic function of the posterior probability of the null, and one could perhaps calibrate or "learn how to interpret P values." This is

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

This content downloaded from 200.130.19.157 on Sat, 28 Mar 2015 14:41:54 UTC
All use subject to JSTOR Terms and Conditions