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.

Table 1. Data, p Values, Posterior Probabilities, and Power at $\theta_1 = .55$ for the Three Surveys

Survey	(a)	(b)	(c)
n	20	200	2,000
$\hat{\theta}$	.750	.575	.523
t	2.03	2.05	2.03
p value	.021	.020	.021
C_n	.408	.816	.976
$\Pr(H_0 \mid t)$	.204	.047	.024
Power(@ 1.645)	.115	.409	.998
Power(@ t)	.057	.262	.993

$\leq \theta \leq .6$. Then t is the usual test statistic, and the p value is $\Phi(-t)$.

A standard calculation yields

$$\Pr(H_0 \mid \hat{\theta}) = \Phi(-C_n t) \tag{1}$$

with

$$C^2{}_n \equiv \tau^2/(\tau^2 + \sigma^2/n) = n/(n + \sigma^2/\tau^2). \tag{2}$$

Note that the probability given in (1) decreases as n increases, in contrast to Jeffreys's formula reported in Table 1 of Berger and Sellke.

The results for the three surveys are reported in Table 1 here.

Survey (a) is far less comforting to Allen than is (b), which is less so than (c). Only for (c), with $C_n = .976$, does $P(H_0 \mid t)$ closely approximate the p value of .021. It is understood in making this assertion that winning and losing are the only items of interest, victory margin being irrelevant (in a real setting, this would be untrue if there were time to influence votes further).

Of course, other results might follow from the same data, but different information. If the election were not expected to be close, for example, if $\tau = .25$ were reasonable, then $C_{20} = .91$ and the p value .021 would be near $\Pr(H_0 \mid t)$ even for $n = 20$. Indeed, this is the Casella–Berger result for the normal distribution setting, that $\Pr(H_0 \mid t)$ diminishes as $\tau \to \infty$ to its minimum $\Phi(-t)$, the p value; check (1) and (2) to see this. Their result is correct, but irrelevant when one knows that τ is bounded above in such a way that C_n is substantially less than unity for all reasonable τ.

The key to understanding these results from any perspective, Bayesian or non-Bayesian, is that the result $\hat{\theta} = .75$ for Survey (a) is not much more likely for the values of θ that one expects to obtain under H_1 than it is if H_0 is true. That is, taking $\theta_1 = .55$ as a typical value for H_1, $\Pr(\hat{\theta} \geq .75 \mid \theta = \theta_1)$ is 5.7% for Survey (a), and it only rises to 12.6% when $\theta_1 = .60$, the largest tenable value for θ. To generalize, and perhaps to explain intuitively when p values fail to reflect probabilities, we note that rare event concepts underlie p value reasoning, but that

if a rare event for H_0 occurs that also is rare for typical H_1 values, it provides little evidence for rejecting H_0 in favor of H_1.

The final two rows of Table 1 provide the powers for the one-tailed tests in each survey at $\theta_1 = .55$, first for test size .05 (rejecting H_0 if $t \geq 1.645$) and in the latter row for test size $\Phi(-t)$, the p value. These power formulas then are $\Phi(\sqrt{n}\delta - 1.645)$ and $\Phi(\sqrt{n}\delta - t)$, respectively, defining $\delta \equiv (\theta_1 - \theta_0)/\sigma$ as the signal-to-noise ratio. Here $\theta_0 = .50$ and $\delta = .1$. We see from Table 1 that

the p value corresponds to $Pr(H_0 \mid t)$ only when good power obtains at typical H_1 parameter values.

I qualify this statement, however, here and in later remarks, by requiring that the parameter space H_1 include the interval between θ_0 and θ_1. Otherwise, in the simple H_0 versus simple H_1 case, for example, there would be excellent power at $\theta_1 = \theta_0 + \delta\sigma$ when δ is large, but at $t = \delta\sqrt{n}/2$, $\hat{\theta} = \theta_0 + \delta\sigma/2$, one has $\Pr(H_0 \mid t) = \frac{1}{2}$, even with a statistically significant test statistic.

Practical statisticians, be they Bayesian or frequentist, have to assess the possible "typical" values θ_1 in H_1 when they design experiments, if only for the purpose of making power calculations to justify the sample size. If we label θ_1 as a typical value when it falls one (prior) standard deviation above the null value θ_0, $\theta_1 = \theta_0 + \tau$, then $C_n^2 = n\delta^2/(1 + n\delta^2)$.

Thus

$$t^* \equiv C_n t \tag{3}$$

is the "corrected" standardized statistic, since then $\Pr(H_0 \mid t) = \Phi(-t^*) = p$ value if t^* had been observed in place of t. Tables of the normal distribution can be applied directly to t^*. In the survey example, taking $t^* = 1.645$ for 5% significance, values of $t = t^*/C_n$ equaling 4.03, 2.01, and 1.69 would be required for $n = 20$, 200, 2,000. Such corrections t^* are in the spirit of the Berger–Sellke rule of thumb for modifying standardized test statistics, but go further because they also incorporate the particular features of each problem.

The essential distinction between the results for two-sided tests and one-sided tests, considered by the authors of these two articles and various others before them, seems not to depend on the number of sides of the test, but on whether all prior probability mass is allowed to slip off to infinity. When that cannot happen, and it automatically cannot in two-sided situations, the p value will tend to be too low. Otherwise, Casella–Berger type results will obtain and p values will be more appropriate. The heuristics of the one-sided survey example are relevant to the Berger–Sellke situation, but the example could easily have been extended to their two-sided situation at the cost of increased complexity.

When significant power is available at reasonable alternatives in H_1, p values will work well. But otherwise they generally overstate evidence. Thus they usually would be reliable for the primary hypotheses in well-designed (for good power) experiments, surveys, and observational studies. But for hypotheses of secondary interest, and

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 $\inf \Pr(H_0 \mid x) \leq p(x)$, 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 $p(x) \leq \Pr(H_0 \mid x)$ 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 $H_0 : \theta \leq 0$ 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

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