Comment

JOHN W. PRATT*

1. BERGER AND SELLKE (AND EDWARDS, LINDMAN, AND SAVAGE)

When I was younger so much younger than today, I never needed anybody's help in any way, least of all the Beatles', and I usually found old fogeys' historical homilies distasteful. As my own fogeyhood impends, I find them just as distasteful, but more salutary. In this vein I must say that, despite the generous references in Berger and Sellke (B&S) and my previous looks at Edwards, Lindman, and Savage (1963) (EL&S), I realized only on recent rereading how much credit is due EL&S for formulating and resolving questions that illuminate the interpretation of P values in testing sharp null hypotheses (and much else). The extent and charm of their penetrating discussion and the progression ordering most of B&S's results are evident in this brief quotation from EL&S (p. 228) on testing the null hypothesis that a normal distribution with known variance has mean $\lambda = 0$.

Lower bounds on L. An alternative when $u(\lambda \mid H_1)$ [the density on H_1] is not diffuse enough to justify stable estimation is to seek bounds on L [the likelihood ratio or Bayes factor in favor of H_0]. Imagine all the density under the alternative hypothesis concentrated at x, the place most favored by the data. The likelihood ratio is then

$$L_{\min} = \frac{\phi(t)}{\phi(0)} = e^{-t^2/2}.$$

This is of course the very smallest likelihood ratio that can be associated with t. Since the alternative hypothesis now has all its density on one side of the null hypothesis, it is perhaps appropriate to compare the outcome of this procedure with the outcome of a one-tailed rather than a two-tailed classical test. At the one-tailed classical .05, .01, and .001 points, $L_{\min}$ is .26, .066, and .0085, respectively. [This essentially covers Th. 1 and Tables 3 and 4 of B&S, in one-tailed form.] Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest. Incidentally, the situation is little different for a two-tailed classical test and a prior distribution for the alternative hypothesis concentrated symmetrically at a pair of points straddling the null value [see B&S, Th. 3 and Tables 2 and 5]. If the prior distribution under the alternative hypothesis is required to be not only symmetric around the null value but also unimodal, which seems very safe for many problems, then the results [B&S, Ths. 5 and 6 and Table 6] are too similar to those obtained later for the smallest possible likelihood ratio obtainable with a symmetrical normal prior density to merit separate presentation here.

After giving results for normal priors (B&S, Th. 8 and Table 7), EL&S "conclude that a t of 2 or 3 may not be evidence against the null hypothesis at all, and seldom if ever justifies much new confidence in the alternative hypothesis" (p. 231) (see B&S, Comment 1).

It is not that B&S claim or sneak off with credit due others. Few are more aboveboard, and I have admired other writing by Berger, in particular his books, for both substance and referencing. But credit slides all too easily onto later authors even when they have no need or desire to steal it. EL&S is still must reading. Do not assume that later publications supersede or subsume it or let its introductory posture or exotic auspices deter you. It is reprinted in at least two books. Only 1% of it is quoted above. The other 99%, though not all so condensed, is also highly rewarding. Some of its subheadings on testing (the topic of half of it) are *Bernoullian example, Upper bounds on L, Haunts of χ^2 and F, Multidimensional normal measurements and a null hypothesis,* and *Some morals about testing sharp null hypotheses.*

B&S's spiraling exposition is helpful the first time around, but afterward I felt a need for more winding up than the graphs of Bayes factors in their Figure 3, even after the trivial but revealing addition of a graph of the comparable frequentist factor $p/(1 - p)$. In the top part of Table 1 here, I have collected and juxtaposed probabilities from B&S's tables (but not the Bayes factors or ratios to pt or pt^2), following A. S. C. Ehrenberg's precepts as best I could. The remaining three lines give $\Pr(H_0 \mid t)$ for a normal prior with variance equal to the sampling variance of the mean (B&S, Table 1 with $n = 1$), and for tight and diffuse priors, which may be viewed as extreme normals (with $n = 0$ and ∞, respectively). Thus the first column shows that the minimum posterior probability for a P value $p = .10$ is .205 when all priors are allowed and increases to .340, .390, and .412 as symmetry, unimodality, and normality restrictions are added. The excess over p and increase with more restrictions on the prior are proportionately even greater at smaller P values. Normality adds little to symmetry, as EL&S observed.

Not to leave well enough alone, I included a "large" t column with B&S's asymptotic formulas and two they happen to omit [where $2.07 = (\pi e/2)^{1/2}$ and $1.77 = \pi^{1/2}$]. They show that the first three are lower bounds for $t > 0$, $t > 2.28$, and $t > 0$, respectively (Theorems 2, 4, 7). The range where the fourth is a lower bound is $t > 2.72$ by my sketchy calculations. (For a normal prior with arbitrary n, the asymptotic formula is $\Pr(H_0 \mid t) = [(n + 1)\pi/2]^{1/2} e^{t^2/2(n+1)} tp$. The range of t where this is a lower bound depends on n. It cannot be a lower bound for all n and t, since it is not a lower bound for $t < 2.72$ in the EL&S worst case $n + 1 = t^2$.)

All the normal results hold for all sample sizes and all prior and sampling variances if n is defined as the ratio of the prior variance to the sampling variance of the mean rather than as the sample size. What I see as "troubling" about the scaling here (see B&S, p. 112) is only the importance of the height of the prior density under H_1 (near $\bar{X}$, say). Such trouble is inevitable in testing sharp null hypotheses, not a deficiency of the prior family. Since n is unrestricted, there is no troubling link between σ and

* John W. Pratt is Professor, Graduate School of Business Administration, Harvard University, Boston, MA 02163. The author is very grateful to Persi Diaconis and Arthur Schleifer, Jr. for helpful comments and to the Associates of the Harvard Business School for research support.

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

Table 1. Comparison of P Values and Minimum $Pr(H_0 \mid x)$ When $\pi_0 = \frac{1}{2}$

t	1.645	1.960	2.576	3.291	Large		
P Value (p)	.10	.05	.01	.001	$\frac{2}{t}\phi(t)$	B&S Tables	B&S Theorems
Priors allowed							
All	.205	.128	.035	.0044	$1.25tp$	3, 4	1, 2
All symmetric	.340	.227	.068	.0088	$2.51tp$	2, 5	3, 4
Symmetric unimodal	.390	.290	.109	.018	t^2p	6	5–7
Symmetric normal	.412	.321	.133	.025	$2.07t^2p$	7	8
Normal var σ^2/n	.42	.35	.21	.086	$1.77e^{t^2/4}tp$	1	
Tight at θ_0	.5	.5	.5	.5			
Diffuse	1	1	1	1			

the prior variance of θ as there is for "conjugate" priors when σ is unknown.

The notion of choosing one or more classical or other insufficient statistics and basing a Bayesian analysis or comparison on them rather than on the whole data set (see B&S, Comment 2) is supported and explored at some length in Pratt (1965, sec. 2).

2. CASELLA AND BERGER (AND PRATT)

In certain one-sided cases, Casella and Berger (C&B, but a different Berger) show that the infimum of $\Pr(H_0 \mid x)$, the posterior probability of H_0, is as small as the P value, p, or smaller. Now a point that permeates EL&S is that, if small, a lower bound is almost useless since it doesn't say you will be anywhere near it. (Hence they seek upper bounds too.) In fact, however, not only is inf $\Pr(H_0 \mid x) \leqq$ or $= p$ but, more to the point, $\Pr(H_0 \mid x)$ itself is close to p in most ordinary one-sided testing problems if n is not small and the prior on θ is not jagged. This is obvious in particular for normal models and hence for procedures concordant with asymptotic likelihood theory. It is also obvious for flat priors in C&B's situation, that of a single observation (or test statistic) x with density known except for location. What C&B add is essentially that, in this situation, $\Pr(H_0 \mid x) < p$ is impossible if the prior is unimodal and the density symmetric with monotone likelihood ratios, but possible in many other cases. Their situation is unfortunately very special. Test statistics, even t and rank statistics, rarely have densities known except for location. Furthermore, for $n > 1$, a regular location family admits a single sufficient statistic only if it is normal with known variance (Kagan, Linnik, and Rao 1973), and otherwise attending to information besides the test statistic can either raise or lower $\Pr(H_0 \mid x)$. So where C&B take us is unclear but not far.

Having done the decent thing and quoted someone else, I will now do the fun thing and quote myself. In Pratt (1965, secs. 7 and 8) I did not merely "state that in the one-sided testing problem the p value can be approximately equal to the posterior probability of H_0" (C&B, p. 106). I *emphasized* the much more important point that it usually *will* be (without claiming novelty even then). I argued both via confidence limits as approximate posterior fractiles and, in location problems, via diffuse priors and independence of θ and $T - \theta$. Among my arguments for confidence limits as approximate posterior fractiles were one's natural reluctance to use them when they are not and asymptotic likelihood theory. I also mentioned Good's elegant argument (1950, 1958). If one-sided reconcilability is as little recognized as C&B suggest, at least I for one tried (both in 1965 and later). But the two-sided discrepancy may get more ink mainly because it is more subtle, surprising, and significant.

As to two-tailed P values, I would have been even more gloomy about the one-dimensional case if I had registered EL&S properly, but what I said in part, partly paraphrased, was "The only widely valid relation between a two-tailed P-value and a posterior probability of natural interest seems to be" that $\frac{1}{2}p$ sometimes has the foregoing one-sided interpretation. Although $1 - p$ "is often approximately the posterior probability that" $0 \leqq \theta \leqq 2\hat{\theta}$, this interval is not of natural interest. Its multidimensional counterpart is "even less so," and indeed depends on irrelevant particulars of the design and test statistic.

> In short, when the null hypothesis $\theta \leqq 0$ is tested against the alternative $\theta > 0$, where θ is one-dimensional and $\theta < 0$ is possible, the P-value is usually approximately the posterior probability that $\theta \leqq 0$. Most other situations where the P-value has a helpful interpretation can be recast in this form. Of course, $\theta \leqq 0$ can be replaced by $\theta \leqq \theta_0$ or $\theta \geqq \theta_0$. And while it is convenient to use P-values in the discussion, those who are interested only in whether or not the results are significant at some preselected level will find similar remarks apply. All the statements about the relation of P-values to posterior probabilities, or lack of it, can be seen easily to hold for a univariate or multivariate normal distribution with known variance or variance matrix. (Pratt 1965, p. 184)

Two technical points. C&B's Lemma 3.1 is an immediate consequence of the fact (subsumed in their proof) that the posterior obtained from a mixture of priors is a mixture of the posteriors obtained from each. The point is more familiar when mixing different models also: the posterior weights are the posterior probabilities of the components, which are of course proportional to their prior probabilities times their predictive densities. B&S (see Th. 3 and its proof) work directly with the Bayes factor and the predictive density, which is equivalent and simpler for the purpose.

C&B's Theorem 3.1 states less than they prove. As it is stated, all but the first sentence of the proof could be replaced by the observation that the inequality follows from Theorem 3.2 (whose proof is independent of Th. 3.1), or directly and easily by considering the uniform prior on $(-k, k)$ as $k \to \infty$ [Eq. (3.5) and the limit calculation at the end of the proof of Th. 3.2].

3. WHAT ABOUT THE PRIORS?

Are the minimizing priors "palatable"? If not, what then? The one-point prior most favorable to H_1 is clearly an exaggeration, but more palatable for one-sided than two-sided alternatives, as EL&S noted. The symmetric two-point prior is still worse for one-sided but somewhat better for two-sided alternatives. EL&S chose accordingly; their remark that one-point priors for one-sided alternatives are "little different" is borne out by halving the P values in B&S's Table 4 and comparing the result with Table 5 (or 2), most easily via the last column unless $p = .05$. All of the minimizing priors depend on the data, an unpalatable feature to most who care at all, and real opinions in one-sided problems would rarely be symmetric or improper. So real prior opinions will often be far from the minimizing opinions, which suggests that real posterior opinions may greatly exceed the lower bounds. This strengthens B&S's main point [because restricting the prior further can only increase the amount by which $\Pr(H_0 \mid x)$ exceeds p in the two-sided case], but points up the weakness of C&B's results in the one-sided case (where matters were already left indeterminate by their argument).

Unfortunately, to discredit a seriously entertained point null hypothesis, one needs something like a lower bound on the prior density in the region of maximum likelihood under the alternative. This appears directly in EL&S but only indirectly in B&S (Comment 3). To my mind it justifies EL&S in being even more cautious in their conclusion (quoted previously) than B&S in Comment 1. Any dimension-reducing hypothesis poses a similar troubling problem. Making such hypotheses approximate makes them more realistic but harder yet to analyze.

4. WHAT'S IT ALL ABOUT?

The broad question under discussion is an important one: what do frequentist inference procedures really accomplish, and what can statisticians of all stripes learn about them by viewing them through Bayesian glasses? The articles here give precise answers to well but narrowly posed subquestions about P values. If you are a Defender of Virtuous Testing or simply a Practical Person, you may feel that the subquestions do not represent the real issues well. But whatever your attitudes or Attitudes, the B&S–EL&S results can hardly comfort you, and I think should disturb you. And even if you can blink them completely—even if you are prepared to disavow any remotely posterior interpretation of P values or visibility through Bayesian glasses—you are not out of the woods. A vast literature discourses on all kinds of problems with hypothesis testing and P values for all kinds of purposes from all kinds of viewpoints: frequentist, Bayesian, logical, practical; for description, inference, decisions, conclusions; preliminary, simultaneous, final; choice of model, estimator, further sampling; and so on. It would be impolite to cite my several nibbles at the subject and invidious to select others, so I will trust the other discussants to suggest its scope. Domains where tests are acceptable may exist, but rejecting Bayesian arguments will not establish or enlarge them.

In summary, I see little major news here beyond what was known by 1963 (EL&S) or obvious by 1965 (Pratt). But every generation must rediscover old truths, and reviving, polishing, and amplifying them and even charting their backwaters are useful. If these articles help the world hear their messages, which I certainly agree with, well and good. If the world is ready for less stylized and precise but all the more disturbing messages about testing, better yet. Regardless, fogeyhood is fun!

ADDITIONAL REFERENCE

Kagan, A. M., Linnik, Yu. V., and Rao, C. R. (1973), *Characterization Problems in Mathematical Statistics* (translated from the Russian by B. Ramachandran), New York: John Wiley.

Comment

I. J. GOOD*

I was interested in both of these articles (which I shall call B&S and C&B) because Bayesian aspects of P values have fascinated me for more than 40 years. The topic will be taken more seriously now that it has hit JASA with two long articles, plus discussion, and the occasion will be all the easier to remember because two Bergers are involved. One result, I hope, will be that the conventional P value of approximately .05, when testing a simple statistical hypothesis H_0, will be correctly interpreted: *not as a good reason for rejecting H_0 but as a reason for obtaining more evidence provided that the original experiment was worth doing in the first place.*

In my opinion P values and Bayes factors are both here to stay, so the relationships between them need to be taken seriously. These relationships form a large part of the main

* I. J. Good is University Distinguished Professor, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. This work was supported in part by National Institutes of Health Grant GM18770.