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.

problem of pure rationality, namely to what extent Bayesian and non-Bayesian methods can be synthesized. (The main problem of *applied* rationality is how to preserve the human species.) My view is that the methods can be synthesized, because, contrary to the opinion of some radical Bayesians, I believe that P values are not entirely without merit. The articles by B&S and C&B contribute to this synthesis, although the *title* of B&S might suggest otherwise.

The relationships between P values and Bayes factors depend on the specific problem, on the background information (some of which is usually vague), on the sample size, on the model assumed, Bayesian or otherwise, and on the questions being asked. B&S and C&B consider distinct questions and, therefore, arrive at distinct solutions. Their problems can be described as significance testing and discrimination, respectively. I think that the article by C&B would have been improved if it had been slightly more friendly to B&S. Television commercials compare burgers, but they do not knock the simple statistical hypothesis. Both articles make useful contributions by careful considerations of inequalities satisfied by Bayes factors. My comments will be partly historical.

Sometimes it is adequate, as in B&S, to define a null hypothesis as $\theta = 0$ or as $|\theta| < \delta$, where δ is small [compare, e.g., Good 1950, p. 91]; sometimes (and this can be regarded as a generalization of the first case) the null hypothesis asserts that $\theta \leqslant 0$ with one or more priors conditional on this inequality; sometimes the initial or prior probability $\Pr(H_0)$ is (approximately) equal to $\frac{1}{2}$ as is usually assumed in both of the articles under discussion and by Jeffreys (1939); sometimes $\Pr(H_0)$ is far from $\frac{1}{2}$ (and of course the posterior probability of H_0 can, therefore, be arbitrarily smaller than a P value); sometimes we prefer to leave the estimation of $\Pr(H_0)$ to posterity and, therefore, try to summarize the evidence from the experimental outcome alone by a P value or by a Bayes factor (or by its logarithm the weight of evidence), both of which have the merit of not depending on $\Pr(H_0)$; and sometimes the priors conditioned on H_0 and on its negation H_1 are reasonably taken as "mirror reflections" in the origin, as is largely assumed by C&B. When testing a treatment that a scientist had previously claimed to be better than a standard one, we are apt to choose H_0 as $\theta = 0$ and H_1 as $\theta > 0$. This model shows more respect to the scientist than if we defined H_0 as $\theta \leqslant 0$ or H_1 as $\theta \neq 0$. Whether he deserves that much respect will again depend on circumstances.

Although the two articles deal with distinct problems, it is possible to produce models that include both problems and intermediate ones. I have worked out one such concrete example that more or less does this and that I shall describe briefly. For more details see Good (in press a). It is a special case of C&B (4.1), but I believe that it is general enough for most purposes.

Let X denote the mean of n random variables, iid, and each $N(\theta, \sigma^2)$, where σ^2 is known or well estimated from the sample. Our aim is to discriminate between $H_0 : \theta \leqslant 0$ and $H_1 : \theta > 0$.

Assume that the prior density of θ given H_i ($i = 0$ or 1) is the folded normal density

$$[(2/\pi)^{1/2}/\tau_i] \exp[-\theta^2/2\tau_i^2)], \tag{1}$$

where $\theta < 0$ if $i = 0$, and $\theta > 0$ if $i = 1$, but with τ_i having the log-Cauchy hyperprior density

$$\psi_i = \frac{\lambda_i}{\pi\tau_i\{\lambda_1^2 + [\log(\tau_i/a_i)]^2\}}. \tag{2}$$

This hyperprior provides a convenient way to give propriety to the familiar improper prior of Jeffreys and Haldane proportional to $1/\tau_i$. The upper and lower quartiles of (2) are $a_i e^{\lambda_i}$ and $a_i e^{-\lambda_i}$, so we can give τ_i a determinate value a_i by letting $\lambda_i \to 0$. In addition, we can determine a_i and λ_i by judging the quartiles.

For this two-level hierarchical Bayesian model we find, after a page of elementary calculus, that the Bayes factor *against H_0 provided by the observation x,* which by *definition* is $O[H_1 \mid (X = x)\ \&\ G]/O(H_1 \mid G)$, is equal to

$$B(H_1 : X = x \mid G) = \Psi_1/\Psi_O, \tag{3}$$

where O denotes odds (also sometimes called an odds ratio), G denotes what was given before X was observed, the colon is read "provided by the information that," the vertical stroke denotes "given" as usual, and

$$\Psi_i = \Psi_i(x, \sigma_n, a_i, \lambda_i) = \int_0^\infty (\sigma_n^2 + \tau^2)^{-1/2} \times \exp\left[\frac{-x^2/2}{\sigma_n^2 + \tau^2}\right]\phi\left[\frac{\varepsilon_i x\tau/\sigma_n}{(\sigma_n^2 + \tau^2)^{1/2}}\right]\psi_i(\tau; a_i, \lambda_i)\, d\tau, \tag{4}$$

where $\varepsilon_0 = 1$, $\varepsilon_1 = -1$, $\sigma_n^2 = \sigma^2/n$ is the variance of X, and ϕ is the error function

$$\phi(y) = (2\pi)^{-1/2} \int_y^\infty e^{-u^2/2}\, du. \tag{5}$$

The integrand in (4) is smooth and not difficult to calculate, so the Bayes factor can be presented as a program with six input parameters, x, σ_n, a_0, a_1, λ_0, and λ_1, and the user can try several priors.

The result contains several interesting special cases, including some results given by B&S and C&B, except that the Bayes factor of B&S will be one half of mine in the appropriate special case. (See my miscellaneous comment 2 below.)

For example, if we take $\lambda_0 = \lambda_1 = 0$, $a_0 = a_1 = \tau$, τ/σ_n large, and $\Pr(H_0) = \frac{1}{2}$, and let H_2 denote the hypothesis that $\theta = 0$, then

$$\Pr(H_0 \mid X = x) \approx \phi(x/\sigma_n) = P,$$

the single-tailed P value corresponding to the "null hypothesis" H_2. Note that H_2 is *not* H_0. We may also describe P as the *maximum* P value over all simple statistical hypotheses of the form $\theta = \theta_0$, where $\theta_0 \leqslant 0$ as in C&B. Because H_2 is not H_0 this case provides only a *partial* reconciliation of Bayesian and Fisherian methods, especially as it is only one of many possible cases, and for this reason I think that C&B have exaggerated. The result certainly does not, and C&B do not claim that it does,

justify the extraordinarily common error, mentioned in both articles, perpetrated by several reputable scientists ("nonspecialists," to quote B&S), of interpreting a P value as $\Pr(H_0 \mid X = x)$ even when H_0 is a point hypothesis. When I mentioned the prevalence of this error to Jim Dickey he pointed out that even Neyman had perpetrated it! [See Good (1984a).] (Most of my citations from now on will be to papers of which I have read every word.)

When $a_0 = 0$, $\lambda_1 = 0$ (so $\tau_1 = \mu_1$), a_1/σ_n is "large," and $x > 2\sigma_n$, we have the situation of B&S (Th. 2, apart from a factor of 2), and the Bayes factor against H_0 is approximately

$$B \approx 2n^{-1/2}(\sigma/\tau_1)e^{s^2/2}, \qquad (s = x/\sigma_n, \text{ the "sigmage"}) \quad (6)$$

$$= \frac{\sigma}{\tau_1 P}\left(\frac{2}{\pi n}\right)^{1/2}\left[\frac{1}{s+}\,\frac{1}{s+}\,\frac{2}{s+}\,\frac{3}{s+}\cdots\right] \quad (7)$$

by Laplace's continued fraction. [Compare Good (1967, p. 410).] Since s is a function of P, it follows that, *for a given value of P,* the Bayes factor against H_0 is proportional to $n^{-1/2}$, and this is usually true when H_0 is a simple statistical hypothesis. This may be called the root n effect and was perhaps first noticed by Jeffreys (1939, pp. 194 and 361–364). For some history of this and allied topics, see Good (1982a).

As a special case of (7) one could append a further column to Table 4 of B&S, giving the values of $O(x)/t$ [or B/t if it is not assumed that $\Pr(H) = \frac{1}{2}$)]. These values would be 1.414, 1.421, 1.391, and 1.350. They are nearly constant because the continued fraction is approximated by $1/s$. This observation is a slight modification of Theorem 2 in B&S.

The root n effect is closely related to the familiar "paradox," mentioned by C&B, that a tail-area pundit can cheat by optional stopping. This possibility is also implicit in Good (1950, p. 96) and was made crystal-clear by reference to the law of the iterated logarithm in Good (1955/1956, p. 13). This form of optional stopping is known as "sampling to a foregone conclusion." To prevent this form of cheating, and to justify to some extent the use of P values as measures of evidence, I proposed "standardizing" a tail-area probability P to sample size 100, by replacing P by $\min(\frac{1}{2}, n^{1/2}P/10)$ (Good 1982b). This proposal is an example of a Bayes/non-Bayes (or Bayes–Fisher) compromise, or "synthesis" as it was called by Good (1957, p. 862) and in lectures at Princeton University in 1955. An example for a multinomial problem was previously given by Good (1950, pp. 95–96). For other examples of the Bayes/non-Bayes synthesis see, for example, Good (in press b).

In most situations that I have seen, where one tests a point null hypothesis, the sample size n lies between 20 and 500, so if we think in terms of $n = 100$, the square root effect will not mislead us by more than a factor of $\sqrt{5}$ in either direction. This explains why I have found that a Bayes factor B' against a point null hypothesis on a given occasion is roughly inversely proportional to P. This leads to the useful harmonic-mean rule of thumb for combining "tests in parallel," that is, tests on the same data (Good 1958, 1984b). This rule of thumb is not precise, but it is much better than the dishonest precise procedure of selecting the test that best supports what you want to believe!

Miscellaneous comments.

1. B&S rightly emphasize the distinction between knowing that $P = P_0$ (or only just less) and knowing only that $P \leqslant P_0$. The latter statement is of course "unfair" to the null hypothesis when P is close to P_0 (Good 1950, p. 94). If a scientist reports only that $P < .05$ we are sometimes left wondering whether $P \approx .049$, in which case the scientist may have been *deliberately* misleading. Such a scientist might have been brought up not to tell fibs, without being told that a flam is usually worse than a fib. Or perhaps he was just brainwashed by an "official" Neyman–Pearson philosophy in an elementary textbook written with the help of a pair of scissors and a pot of glue and more dogmatic than either Neyman or Egon Pearson were. If Neyman had been dogmatic he would not have made the "nonspecialist's error," or error of the third kind, mentioned previously.

2. In the past, and frequently in conversation, I have used a rough rule that a P value of .05 is worth a Bayes factor of only about 4 when testing a simple statistical hypothesis (e.g., Good 1950, p. 94; 1983, p. 51). B&S get about half this value because they use a prior symmetric about $\theta = 0$ given H_1, whereas my rule is intended more for the case in which H_1 asserts that $\theta > 0$.

3. The topic of max factors, mentioned by B&S, without the cosmetic name, was also discussed in Good (1950, p. 91) as applied to multinomials, which of course includes binomials, and where the maximum weight of evidence (log-factor) is related to the chi-squared test. In the binomial case, the approximation given for the maximum weight of evidence (in "natural bans") again H_0 naturally agrees with the result $\frac{1}{2}t^2$ cited in Example 1 of B&S. Although in multivariate problems the max factor is much too large, the relationship to χ^2 shows the relevance to an aspect of the philosophy of the Bayes/non-Bayes or Bayes–Fisher synthesis, namely that even a poor Bayesian model can lead to a sensible non-Bayesian criterion (a point that I have made on several other occasions).

Sometimes a multivariate test can be reduced to a univariate one. B&S mention an example, and another example is that of a max factor that is useful because the maximization is over a *single* hyperparameter as in the mixed Dirichlet hierarchical Bayes approach to multinomials and contingency tables (e.g., Good 1976, p. 1170; Good and Crook 1974, p. 714).

4. In their concluding comments B&S state that when considering a simple statistical hypothesis H_0, by and large 2σ is weak evidence against H_0, 3σ is "significant," and so on. These conclusions agree roughly with Good (1957, p. 863), where I judged that the Bayes factor in favor of H_0 usually lies within a factor of 3 of $10P$. (This can break down if $P < 1/10{,}000$ and for very large sample sizes.)

5. The references in B&S cover much of the literature, and this will presumably be more true when the comments

are included. To aid in making the bibliography more complete I exercise the rights of a senior citizen and list 28 additional relevant publications of which I have read every word (10 of them are in the conscientious reference list of B&S): (a) items C73, C140, C144, C199, C200, C201, C209, C213, C214, and C217 in *Journal of Statistical Computation and Simulation* (1984); (b) Items 13 (pp. 91–96), 82, 127 (pp. 127–128), 174, 398 (p. 35), 416, 547, 603B (p. 61), 862, 1234 (pp. 140–143), 1278 (regarding Bernardo), 1320–C73, 1396 (pp. 342–343), 1444, and 1475–C144 in the bibliography (pp. 251–266) in Good (1983); (c) Good (1955/1956, p. 13; 1981; 1983, indexes; 1986; in press a,b). To these may be added the thesis of my student Rogers (1974) and a further reference relevant to C&B, Thatcher (1964).

ADDITIONAL REFERENCES

Good, I. J. (1955/1956), Discussion of "Chance and Control: Some Implications of Randomization," by G. S. Brown, in *Information Theory, Third London Symposium 1955,* London: Butterworth's, pp. 13–14.

——— (1957), "Saddle-Point Methods for the Multinomial Distribution," *Annals of Mathematical Statistics,* 28, 861–881.

——— (1976), "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *The Annals of Statistics,* 4, 1159–1189.

——— (1981), Discussion of "Posterior Odds Ratio for Selected Regression Hypotheses," by A. Zellner and A. Siow, *Trabajos de Estadistica y de Investigacion Operativa,* 32, No. 3, 149–150.

——— (1982a), Comment on "Lindley's Paradox," by Glenn Shafer, *Journal of the American Statistical Association,* 77, 342–344.

——— (1982b), "Standardized Tail-Area Probabilities" (C140), *Journal of Statistical Computation and Simulation,* 16, 65–66.

——— (1984a), "An Error by Neyman Noticed by Dickey" (C209), in "Comments, Conjectures, and Conclusions," *Journal of Statistical Computation and Simulation,* 20, 159–160.

——— (1984b), "A Sharpening of the Harmonic-Mean Rule of Thumb for Combining Tests 'in Parallel' " (C213), *Journal of Statistical Computation and Simulation,* 20, 173–176.

——— (in press a), "A Flexible Bayesian Model for Comparing Two Treatments," C272, *Journal of Statistical Computation and Simulation,* 26.

——— (in press b), "Scientific Method and Statistics," in *Encyclopedia of Statistical Science* (Vol. 8), eds. S. Kotz and N. L. Johnson, New York: John Wiley.

Good, I. J., and Crook, J. F. (1974), "The Bayes/Non-Bayes Compromise and the Multinomial Distribution," *Journal of the American Statistical Association,* 69, 711–720.

Jeffreys, H. (1939), *Theory of Probability* (1st ed.), Oxford, U.K.: Clarendon Press.

Rogers, J. M. (1974), "Some Examples of Compromises Between Bayesian and Non-Bayesian Statistical Methods," unpublished doctoral thesis, Virginia Polytechnic Institute and State University, Dept. of Statistics.

Thatcher, A. R. (1964), "Relationships Between Bayesian and Confidence Limits for Predictions" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 26, 176–192.

Comment

DAVID V. HINKLEY*

The authors have added an impressive array of technical results to the main body of work on this subject by Jeffreys, Lindley, and others. The sense of surprise in the first article suggests that statistical education is not as eclectic as one might wish. In my brief comments I should like to mention some of the general issues that should be considered in any broad discussion of significance tests.

First, the interpretation of P value as an error rate is unambiguously objective and does not in any way reflect the prior credibility of the null hypothesis. Rules of thumb aimed at calibrating P values to make them work like posterior probabilities cannot reflect the broad range of practical possibilities: in many situations the null hypothesis will be thought not to be true.

One area where null hypotheses have quite high prior probabilities is model checking, including both goodness-of-fit testing and diagnostic testing. Here specific alternative hypotheses may not be well formulated, and significance test P values provide one convenient way to put useful measures on a standard scale.

Rather different is the problem of choosing between two, or a few, separate families of models. Here the symmetric roles of the hypotheses seem to me to make significance testing very artificial. It would be better to adopt fair empirical comparisons, using cross-validation or bootstrap methods, or a full-fledged Bayesian calculation. The latter requires careful choice of prior distributions within each model to avoid inconsistencies.

Significance tests will sometimes be used for a nuisance factor, preliminary to the main test, as with the initial test for a cross-over effect in a comparative trial with cross-over design. Racine, Grieve, Fluhler, and Smith (1986) recently demonstrated the clear merits of a Bayesian approach in this context. If significance tests are to be useful, then they should have validity independent of the values of identifiable nuisance factors.

In general, for problems where the usual null hypothesis defines a special value for a parameter, surely it would be more informative to give a confidence range for that parameter. Note that some significance tests are not compatible with efficient confidence statements, simply

*David V. Hinkley is Professor, Department of Mathematics, University of Texas, Austin, TX 78712.

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