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

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

possible, however, only if the calibration is fairly simple and *constant*. In our article we mentioned one well-known source of nonconstancy in interpretation of the P value: as the sample size increases in testing precise hypotheses, a given P value provides less and less real evidence against the null. One could perhaps argue that a different calibration can be found for each sample size. But now Casella and Berger have also demonstrated that one must calibrate by the nature of the problem. For one-sided testing, a P value is often roughly equivalent to evidence against H_0, whereas for testing a precise hypothesis a P value must typically be multiplied by a factor of 10 or more to yield the same evidential interpretation. And these are *not* the only two possibilities. Indeed, suppose that the null hypothesis is an interval of the form $H_0 : |\theta - \theta_0| \leq C$. If C is near 0, one is effectively in the point null situation, and as C gets large the situation becomes similar to one-sided testing. For C in between, there is a continuum of different possible "calibrations."

Although somewhat less important than the sample size and C, the dimension of the problem and the distribution being considered can also necessitate different calibrations between P values and "evidence against H_0." The bottom line is simple: the concept of a P value is faulty, in that it does not have a reasonable direct interpretation as to evidence against H_0 over the spectrum of testing problems. It may be useful to identify when P values are (and are not) sensible measures of evidence, so as to allow reappraisal of those scientific results that have been based on P values, but the future of the concept in statistics is highly questionable.

Another issue raised in the article of Casella and Berger has to do with the validity of precise hypothesis testing. It is implied in Section 1 that one-sided tests are more useful in practice, and in Section 4 that placing mass near a point can be considered as "biasing the result in favor of H_0"; the practical import of our results is thus questioned. This issue is complicated by the fact that, in practice, many testing problems are erroneously formulated as tests of point null hypotheses. There is undeniably a huge number of such tests performed, but how many should be so formulated?

One answer to this objection is simply to note that we have little professional control over misformulations in statistics; we do, however, have some control over the statistical analysis performed for a given formulation. It is awkward to argue that a bad analysis of a given formulation is okay because the formulation is often wrong.

At a deeper level, it is possible even to argue the other way on the question of proper formulations of testing; one can argue that it is actually precise nulls that encompass the majority of "true" testing problems. This argument notes that most one-sided testing problems have to do with things like deciding whether a treatment has a positive or negative effect, or which of two treatments is best. The point is that, in such problems, what is typically really desired is an evaluation of how *large* the effect is or how *much* better one treatment is than another. Such problems are more naturally formulated as estimation or decision problems, and the appropriateness of testing is then debatable.

Precise hypotheses, on the other hand, ideally relate to, say, some precise theory being tested. Of primary interest is whether the theory is right or wrong; the amount by which it is wrong may be of interest in developing alternative theories, but the initial question of interest is that modeled by the precise hypothesis test.

In such problems the key fact is that there *is* real belief that the null hypothesis could be approximately true. If I am an experimenter conducting a test that will show, hopefully, that vitamin C has a beneficial effect on the common cold, I had better officially entertain the hypothesis that its effect is essentially negligible. In other words, I should not take the prior mass assigned to "no positive effect" and spread it out equally over all $\theta \leq 0$; this does not correspond to the reality that most people may be quite ready to believe that vitamin C is not harmful, yet give substantial weight to a belief in no or little effect. Such situations require substantial prior mass near 0.

We present the previous argument about what is "practical hypothesis testing" only halfheartedly. The huge variety of applications in which P values are used (see Cox 1977) makes questionable any claim that only "one type" of situation need be considered from a practical perspective. Whether most situations are one-sided, have a precise null hypothesis, or are really decision problems is irrelevant; our basic statistical theory should handle all.

2. REPLY TO HINKLEY

Hinkley defends the P value as an "unambiguously objective error rate." The use of the term "error rate" suggests that the frequentist justifications, such as they are, for confidence intervals and fixed α-level hypothesis tests carry over to P values. This is not true. Hinkley's interpretation of the P value as an error rate is presumably as follows: the P value is the Type I error rate that would result if this observed P value were used as the critical significance level in a long sequence of hypothesis tests [see Cox and Hinkley (1974, p. 66): "Hence [the P value] is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as being just decisive against H_0."] This hypothetical error rate does not conform to the usual classical notion of "repeated-use" error rate, since the P value is determined only once in this sequence of tests. The frequentist justifications of significance tests and confidence intervals are in terms of how these procedures perform when used repeatedly.

Can P values be justified on the basis of how they perform in repeated use? We doubt it. For one thing, how would one measure the performance of P values? With significance tests and confidence intervals, they are either right or wrong, so it is possible to talk about error rates. If one introduces a decision rule into the situation by saying that H_0 is rejected when the P value $\leq .05$, then of course the classical error rate is .05, but the expected P value given rejection is .025, an average understatement of the error rate by a factor of two.

In the absence of an unambiguous interpretation of P values as a repeated-use error rate, we have most frequently heard P values defended as a measure of the evidence against H_0, via an "either H_0 is true or a rare event has occurred" argument. It is for this reason that we concentrated on evaluating P values in terms of whether they really are effective in conveying information about the strength of the evidence against H_0. We acknowledge the difficulty in defining "evidence" in an absolute (non-Bayesian) sense, and for this reason we considered a variety of notions of evidence in the article, including lower bounds on the Bayes factor (or weighted likelihood ratio). Indeed, the lower bound on the Bayes factor strikes us as having a true claim to being "unambiguously objective," since it depends on no prior inputs at all (Th. 1) or only on a symmetry assumption (Th. 3) and yet relates to a valid (conditional) measure of evidence.

We indicated in Comment 2 that the results can be extended to goodness-of-fit testing and yield much the same conclusions, even when the alternative hypotheses are not well formulated. One can find lower bonds over essentially arbitrary alternatives within the chi-squared testing framework. Thus, whether or not the P value can really be considered as a standard scale, its interpretation in terms of evidence against H_0 should be sharply qualified.

We would disagree with the idea that usual confidence ranges for a parameter are more informative than posterior probabilities of hypotheses, when the null hypothesis defines a special value for a parameter. As an example, the density (on $\mathbf{R}^1$)

$$f(x|\theta) = (1 + \varepsilon) - 4\varepsilon|x - \theta|, \qquad \text{for } |x - \theta| \leq \tfrac{1}{2},$$

will yield, as a usual 95% confidence set for small ε,

$$C(x) = (x - .475, x + .475);$$

but if $\theta = 0$ is a special value and $x = .48$ is observed, we would be loathe to reject $H_0 : \theta = 0$, since

$$f(.48|0)/\sup_{\theta \in C(.48)} f(.48|\theta) \geq (1 - \varepsilon)/(1 + \varepsilon).$$

The point is that a special parameter value outside a confidence set can have virtually the same likelihood as any parameter value inside a confidence set, and we would then argue that the data do not indicate rejection of the special parameter value. This phenomenon also occurs in the normal testing problem we discuss, though to a lesser degree.

We are wholeheartedly in agreement that proper conditioning must be employed. To us, however, this is even more important in testing than with confidence sets. We feel that refusing to "condition" on the actual data x, and instead using the set A of "as or more extreme" values, causes more harm in statistical practice than other failures to condition.

3. REPLY TO VARDEMAN

Our major disagreement seems to center again on the issue of concentrating prior mass near θ_0. We argued previously that (a) in examples such as the "vitamin C" example, one often does have mass near θ_0, and (b) even if H_0 is a fair-sized interval, the contradiction occurs (the agreement of posterior probabilities with P values only occurring in the limiting case in which H_0 is a very large interval with prior mass "uniformly" distributed over it).

Perhaps less controversy would have ensued if we had used Bayes factors or weighted likelihood ratios as our central measure. The argument then avoids the loaded issue of "prior beliefs" and simply says "how does the support of the data for H_0, given by the likelihood $f(x \mid \theta_0)$, compare with the support of the data for H_1, given by some average of $f(x \mid \theta)$ over θ in H_1." This is the Bayes factor, with g being the averaging measure on H_1, and the various theorems find bounds on the Bayes factor over g. If θ_0 has no distinction, as in the scenario of Casella and Berger, one probably does not care if $f(x \mid \theta_0)$ is a substantial fraction of the weighted likelihood of H_1; on the other hand, if θ_0 has the distinction of being a particular value for which it is desired to assess the evidence for or against, it is hard to ignore a comparatively large value of $f(x \mid \theta_0)$. We chose not to emphasize this "likelihood" argument, because we have found that the interpretation of observed likelihood ratios as direct evidence (and not just as inputs into a classical test) is less familiar to many classical statisticians than is the use of posterior probabilities as evidence.

This also relates to the issue of our agreed-upon discomfort at replacing $t = 1.4$ by the event $[|t| \geq 1.4]$. In the normal case (and most others), $f(1.4 \mid \theta_0)$ is a substantial fraction of any reasonable average of $f(1.4 \mid \theta)$ over H_1. On the other hand, $\Pr([|t| \geq 1.4] \mid \theta_0)$ is much smaller than reasonable averages of $\Pr([|t| \geq 1.4] \mid \theta)$ over H_1. Thus, by likelihood reasoning, there is also a great difference between knowing precisely that $t = 1.4$ and knowing only that $|t| \geq 1.4$; the latter would yield much greater evidence against H_0.

Another illustration of the conditioning aspect of the problem is described in our story about the "astronomer" in Section 1. We would really like to see an explanation, written for this astronomer, as to why he should believe that $t = 1.96$ is substantial evidence against H_0. The general point is that any method of conditionally measuring evidence that we have considered indicates that the replacement of $t = 1.4$ by $[|t| \geq 1.4]$ is the source of the huge discrepancies; and the replacement has no real justification except that of "convenience." One of the purposes of this article was to indicate a common statistical situation in which it is essential to condition properly, feeling that the issue of conditioning *is* one of the deepest and most important issues in statistics.

We applaud Vardeman's leanings toward decision-theoretic formulations, though we have argued that one should not completely abandon the possibility of stating how much the data support a special value θ_0. We also are not particularly at ease with the use of words like "objective," but we use them out of a certain defensive posture. Many statisticians feel that it is possible and essential to be objective; whether or not this really is possible, we

would argue that the closest one can come to objectivity is through the types of conditional analyses we have discussed. (See Comment 3 for our views concerning the actual possibility of objectivity.)

4. REPLY TO DICKEY

The observation that n in Table 1 can, in general, be replaced by the ratio of the prior and sampling variances is a useful fact (pointed out also by Pratt). It is interesting that the accuracy of the point null formulation (i.e., the appropriateness of the approximation of a realistic small interval null by a point) depends on $\sigma/\sqrt{n}$ but not on τ^2; thus if τ^2 is indeed larger than σ^2, one can move to the right in the table without increased worry concerning the soundness of the formulation.

The asymptotic t arguments are given for completeness, but it is true that the asymptotics take effect for t too large to be of much interest. We agree with all other comments, except that the equating of a P value with a tom-tom strikes us as somewhat overly positive.

5. REPLY TO PRATT

We are in complete agreement that Edwards, Lindman, and Savage (1963) (EL&S) contained the essence of our article. Indeed, had EL&S not been so mysteriously ignored for so long, our contribution would have been mainly a presentation of Theorem 5, its ramifications, and the results in Section 4. Because very few people we talked to were aware of the results in EL&S, however, a general review seemed to be in order. We feel that the result of Theorem 5 is a substantive advance for two reasons. First, although the results for G_{US} are not greatly different from those for G_N, this is not apparent a priori; non-Bayesians tend to be very wary of a result established for only normal priors, so verifying that the same answer holds qualitatively for all unimodal symmetric priors can substantially enhance the impact of the basic phenomenon. Second, the techniques for working with large classes, such as G_{US}, are important in general Bayesian sensitivity studies, and we hoped that the application here would indicate the possibilities and kindle interest. Finally, the result on interval hypotheses in Section 4 is valuable for both sociological and scientific purposes.

Pratt's Table 1 and the subsequent comments and insights are all of value. We agree with his later comment that our Comment 1 is probably not cautious enough; it was given with the simple hope that a not-too-terrible rule of thumb might be able to drive out a terrible rule of thumb.

6. REPLY TO GOOD

There is virtually nothing in this interesting set of comments with which we disagree. We would probably have to align ourselves with the radical Bayesians, however, in that we remain unconvinced that P values have any merit. The number of "rules of thumb" that have to be learned to "calibrate" properly P values in the various possible testing situations is so large that it strikes us as simply unwieldy to continue to use them. Why not just shift over to Bayes factors (or bounds on the Bayes factors)? We would agree that often (though not always) a P value of .05 is an indication that more evidence should be obtained.

We thank Good for the additional references; we tried, but knew we must have missed some.

7. REPLY TO MORRIS

Morris raises a number of interesting issues that bear on the comparison of the one-sided and precise null testing situations. For ease in discussion, it is helpful to consider a precise null version of the example of Morris.

Example. Consider a paired comparison experiment in which two new treatments will be screened. The outcome for each subject pair is a 0 or 1, depending on which treatment is judged to be superior. Let θ denote the probability of obtaining a 1, and let n denote the number of (independent) pairs in the experiment. These are two new treatments, and it is judged that there is a substantial probability ($\frac{1}{2}$, say) that they are both ineffective, which would correspond to a θ very near $\frac{1}{2}$. All past experiments with similar treatments have indicated that, when there are treatment effects, θ ranges between .4 and .6. Indeed (as in the Morris example), suppose that we view it reasonable to model this θ, a priori (conditional on there being treatment effects), as having an $\mathcal{N}(\frac{1}{2}, (.05)^2)$ distribution. Assuming that the normal approximation for $\hat{\theta}$ is valid, the entire model above falls within the framework of our article, with $X = \hat{\theta} \sim \mathcal{N}(\theta, .25/n)$, the desire to test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta \neq \frac{1}{2}$, $\pi_0 = \frac{1}{2}$, and $g(\theta)$ being the $\mathcal{N}(\frac{1}{2}, (.05)^2)$ density.

The difference between this problem and that of Morris is, of course, that there *is* substantial reason to suspect $\theta = \frac{1}{2}$; in a voting situation there is no reason to single our $\theta = \frac{1}{2}$ as deserving positive prior mass. (We implicitly assume that n is not enormous; the real hypothesis of "no treatment effects" would be accurately modeled as $H_0 : |\theta - \frac{1}{2}| \leq \varepsilon$, and if n is enormous it can be inaccurate to approximate this by $H_0 : \theta = \frac{1}{2}$.)

By using an easy modification of formula (1.1), we can calculate the posterior probability of H_0 for each of the situations in Table 1 of Morris. The results for $n = 20$, $n = 200$, and $n = 2{,}000$, respectively, are .436, .302, and .387; compare these with the posterior probabilities found by Morris of .204, .047, and .024, respectively. Note, in particular, the huge difference for $n = 2{,}000$.

The example here makes clear that the insightful comments of Morris, although valid for the situation in which no special mass is to be assigned to a point θ_0, need not be valid for the precise null situation. For instance, the comment "the P value corresponds to $\Pr(H_0 \mid t)$ only when good power obtains at typical H_1 parameter values" may be valid for nonprecise nulls but is false for precise nulls; the powers at $\theta = .55$ for our example are very near 1

when $n = 2{,}000$, yet the P value differs drastically from the posterior probability of H_0.

The necessary distinction between precise nulls and imprecise nulls only reinforces the exhortation (with which we completely agree), in the last paragraph of Morris's comment, to the effect that it is crucial for all statisticians and scientists using P values to learn exactly what P values do and do not convey about the evidence against H_0 in the wide variety of testing problems to which they are applied.

ADDITIONAL REFERENCES

Cox, D. R. (1977), "The Role of Significance Tests," *Scandinavian Journal of Statistics,* 4, 49–70.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics,* London: Chapman & Hall.