the technique of testing hypotheses is vastly overrated in statistics as a method. It isn't so much that the classical methods give the wrong answers, as Berger and Delampady correctly show, as it is that I find the problem ill-suited to help me do statistics better. Thus, I find myself in agreement with Berger and Delampady that "when testing precise hypotheses, formal use of P-values should be abandoned." On the other hand, I do not expect to test a precise hypothesis as a serious statistical calculation.

ADDITIONAL REFERENCES

KADANE, J. B., LEWIS, G. and RAMAGE, J. (1969). Horvath's theory of participation in group discussion. *Sociometry* **32** 348–361.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

Rejoinder

James O. Berger and Mohan Delampady

We are grateful to the discussants for their comments. All raise interesting issues that are highly deserving of discussion. As usual, we will focus on disagreements in our rejoinder.

REPLY TO COX

Professor Cox questions our argument that P-values do not have a valid frequentist interpretation, stating that the "hypothetical long-run frequency interpretation of a significance level seems totally clear and unambiguous." Over many years of trying to understand what makes a valid frequentist interpretation, we have come to agree with Neyman's view that one must have a stated accuracy criterion, a stated procedure and determine the expected accuracy of the procedure in repeated use; thus, an $\alpha = .05$ level test will indeed reject true nulls only 5% of the time in repeated use. A P-value has no such *real* frequentist interpretation. It has various pseudofrequentist interpretations (cf. Cox and Hinkley, 1974), but these are somewhat contorted so that their impact, or persuasiveness, is much less than that of the *real* frequentist justification. Also, a thorough study of our Example 6 is, we feel, very important in understanding the role of frequentism here.

The reaction of Cox to our claim, that ". . . inclusion of all data 'more extreme' than x_0 is a curious step and one we have seen no remotely convincing justification for," is to say that he finds the reasoning clear and precise and at least sometimes relevant. He, of course, is well aware of the many examples in statistics (some due to Cox himself) where inclusion of "other data" in the calculation leads to nonsense. We submit that this is one of those situations, and indeed can marshall (following Jeffreys) purely intuitive arguments against including more extreme data: is it really fair to H_0 to hurl against it not just the (mild) evidence x_0, but also all the much stronger "extreme" values, when these extreme values *did not occur*?

We, for the most part, agree with the remaining comments of Cox. Our statement that "formal use of P-values should be abandoned" was directed to the formal use of P-values in providing quantitative measures of doubt of H_0. At the beginning of Section 5 we agreed that the informal use of P-values "as a general warning that something is wrong (or not) . . . " (to use Cox's phrase) is perhaps reasonable; this informal use in data analysis may well justify the teaching and consideration of P-values.

In regard to "sensible uses of P-values," it is worth considering an earlier comment of Cox to the effect that for "dividing hypotheses . . . the apparent disagreements between different approaches are normally minor." We used to think this, but the discussion of Carl Morris to Berger and Sellke (1987) shows that such may well not be so.

Finally, our response to Cox's Rejoinder 8 or 4′ is what would be expected of Bayesians: We feel that using the Bayesian paradigm will give misleading answers less often than use of alternative paradigms.

REPLY TO EATON

We agree with just about everything in Professor Eaton's discussion, leaving us little to do but applaud the further insights provided. The objectivity issue is indeed a fundamental concern. Eaton argues that objectivity is a vague, ill-defined concept, and may not exist. We agree; indeed, one of the major purposes of the paper was to show that Opinion 2 in the introduction is wrong. Testing a precise hypothesis is a situation in which there is *clearly* no objective Bayesian analysis and, by implication, no sensible objective analysis whatsover. In other problems, arguments about whether noninformative priors are, or are not, objective tend to be inconclusive, but here there simply is *no* prior that can even be called noninformative.

Although the precise hypothesis testing scenario was used to demonstrate that objectivity is at least

sometimes unobtainable, there are compelling practical reasons to have, on the shelf (as Zellner says), procedures that have the appearance of objectivity. History has shown us that very many users of statistics will only use off-the-shelf procedures that do not require overt subjective inputs, and we feel that it is important to provide a Bayesian version of this shelf. The use of the phrase "conventional procedures," rather than "objective procedures," to describe the shelf is probably less objectionable and misleading. Of course, we are not denying the force of Eaton's argument; there is no good substitute for thinking about the problem and the subjective inputs that are required. But among the substitutes for thinking, we would argue that use of the conventional Bayesian tests is vastly superior to use of P-values.

REPLY TO ZELLNER

We thank Professor Zellner for the kind words, numerous additional references, and discussion of relevant work. We had purposely avoided extensive discussion of the operational side of Bayesian testing, so as to keep the scope of our paper reasonable. From Zellner's comments it is clear that this was an error; many additional fundamental insights are available in this literature. We organize our responses here according to the numbering scheme used by Zellner.

2. General Points

Point 1. Zellner defends exact point nulls (in interesting contrast to Kadane) and provides an example from physics where θ_0 is a specific physical constant. We certainly did not mean to preclude exact point nulls; our purpose was mainly to show that exactness is not a necessary requirement for our results.

It is, however, of interest to briefly discuss the realism of exact nulls, H_0: $\theta = \theta_0$. There is little question that such nulls are reasonable conceptually, but are they reasonable in practice? The problem is that, in real experiments, there is virtually always some type of bias. Thus, we don't observe $X_i \sim N(\theta, \sigma^2)$, say, but instead observe $X_i \sim N(\theta + \varepsilon, \sigma^2)$, where ε is perhaps very small, but unknown. All we can ever actually test, using such data, are hypotheses like H_0: mean of $X_i = \theta_0$ or, equivalently, H_0: $\theta = \theta_0 - \varepsilon$, i.e., that θ is close to θ_0. Whether or not experiments exist with exactly zero bias, allowing tests of exact point nulls, is not really worth discussing, because our results indicate that one can assume that the null is exact if the bias is small enough.

Point 2. This is an excellent point. Considering H_1: $\theta = 0$ vs. H_2: $\theta < 0$ vs. H_3: $\theta > 0$ is often very desirable, and synthesizes many of the ideas. It is a strength of the Bayesian approach that it can as easily deal with multiple hypotheses like these, as with two hypotheses.

Point 3. We agree with this point, except for the implication that on-the-shelf testing procedures correspond to little prior information. They correspond to the specific prior belief (when g is Cauchy(θ_0, σ^2)) that, under H_1, θ has half interquartile range σ. Our point here was that there is *no* sensible prior that truly corresponds to "little previous information." Again, however, we agree to the need for off-the-shelf procedures, and support the use of Jeffreys's, Zellner's and others' "conventional" priors.

Points 4 *and* 5. We agree.

3. Technical Points

Point 1. We do not fully understand this comment. We attempted to indicate when a small interval null could be approximated by an exact point null, because the prior elicitation process can be greatly simplified in this case. Determining ε and the shape of the prior spike in the null interval is very hard, and characterizing situations where difficult prior assessments can be avoided (because of robustness) is surely not a waste of time.

Point 2. Our own preference is indeed for the Cauchy prior, rather than the normal prior, for robustness reasons. The difference between the normal and Cauchy priors here is really not all that great, however, especially if one properly scales the priors. The key feature in scaling is to equate the height of the prior densities as $\theta \to \theta_0$, this height being the dominant term in any expansion of the Bayes factor. Matching heights would result, as the normal analogue of the Cauchy (θ_0, σ^2) density, in a $N(\theta_0, \pi\sigma^2/2)$ density (which also more nearly matches the quartiles of the Cauchy density). The Bayes factor for this normal analogue is

$$BF = \left(1 + \frac{1}{2}\, n\pi\right)^{1/2} \exp\left\{-\frac{1}{2}\, t^2 \Big/ \left(1 + \frac{2}{n\pi}\right)\right\},$$

which will give very similar answers to Jeffreys (e.g., for $n = 100$ and $t = 1.96$, $BF = 1.86$ compared to Jeffreys's $BF = 1.91$), except possibly for small n and large t (where there is little question as to what the conclusion should be).

Point 3. We did not attempt to compare the various conventional priors that have been put forth, although this comment makes us wish that we had! For our reaction to "little information," see our response to Section II, Point 3.

Point 4. We entirely agree that the class (10) is not suitable for actual use in arriving at *an* answer, but that was not its purpose. Its purpose was to unequivocably show the inappropriateness of the P-value, and to indicate a reasonable lower bound on

the Bayes factor. As we discussed in Section 5, however, this lower bound does not generally provide a specific enough answer for a Bayesian; one must consider specific prior information for the problem at hand, as Zellner suggests.

4. Concluding Remarks

We agree with all comments in this section. In our defense, we did not set out to provide a careful review or a careful presentation of how a Bayesian should handle hypothesis testing. The focus of our paper was on a discussion of certain not-well-understood or controversial issues. A general review of Bayesian hypothesis testing would have provided much more background on Jeffreys's (and Zellner's) work (and would have also been another hundred pages). For those who feel ready to try to do it the Bayesian way, our paper is not a substitute for reading the Bayesian literature.

REPLY TO BAYARRI

Professor Bayarri raises several extremely interesting issues. The first is that the assumptions in Section 2.2, under which we established that an interval null can be approximated by a point null, cannot be satisfied by many common priors, such as conjugate priors. This is essentially correct. Indeed, about the only easy way to construct an overall prior $\pi(\theta)$, which satisfies the conditions, is to choose it to be a *mixture* of a concentrated distribution (a spike) in the interval, and a more diffuse distribution outside the interval. For precise hypothesis testing, however, we would argue that such priors are inherent to the problem, and that priors such as conjugate priors are unnatural. Thus, for H_0: vitamin C has no effect on the common cold, prior elicitation will typically involve two distinct thought processes. The first recognizes that there will be a concentration of mass near "no effect" (the spike), and the second recognizes that vitamin C could have an effect, but if so the effect could be quite substantial (i.e., $\pi_1(\theta)$ will not itself be concentrated near zero). This kind of prior information virtually requires a mixture distribution for effective modeling. In essence, we are *defining* precise hypothesis testing as that class of testing situations in which the prior is of the above type.

It is actually technically possible to apply Theorem 1 (the approximation theorem) to more typical prior distributions, such as conjugate priors, as long as the distributions are fairly diffuse; it will still be the case that B is approximately equal to $\hat{B}$. For such distributions, the prior probability of H_0 will tend to be small (see also our "Reply to Casella and Berger"), but approximation by a point null remains possible. Thus, the only situations really excluded from consideration are those in which the prior mass assigned to H_1 is mostly concentrated very near H_0. Such situations are rare.

Bayarri next turns to the class of important testing problems (also discussed by Casella and Berger) in which θ_0 has no special prior believability, but is important for reasons of simplification or utility. We agree that the important feature of such problems is usually that of determining whether or not θ is close enough to θ_0 to allow use of the simplifying θ_0, and that the type of decision-theoretic technique discussed by Bayarri might well be very useful in this regard. Bayes factors or posterior probabilities of hypotheses will often be of little use for this type of problem (although note that if θ_0, in addition, has approximate prior believability then, as discussed in our Rejoinder 7, the results of the paper do become relevant to a decision-theoretic analysis).

REPLY TO CASELLA AND BERGER

General Remarks

The main thesis of Casella and Berger's interestingly provocative discussion strikes us as a prime example of what I. J. Good calls the non-Bayesian proclivity to SUTC (sweep-under-the-carpet) subjective input. Let us recast what Casella and Berger seem to be saying in terms of a hypothetical Bayesian and non-Bayesian report from an experiment.

Non-Bayesian. I have determined that the P-value is 0.05, and so there is significant evidence that H_0 is wrong.

Bayesian. My prior probability of H_0 is only 0.1, and the Bayes factor against H_0 is ½. Hence the posterior probability of H_0 is about 0.05, which I feel is significant evidence against H_0. (Note, from Table 1 in the paper, that if the P-value is 0.05 and n is 10 or less, then the Bayes factor will be about ½.)

The Bayesian feels confident that H_0 is wrong, but clearly states that, of the total evidence against H_0, only a factor of ½ is due to the data, while a factor of 1/10 (clearly the most influential component of the evidence) is due to prior opinion. Casella and Berger, on the other hand, seem to be arguing that it is okay to report a P-value of 0.05 in situations such as this, *and* to interpret it as significant evidence against H_0, because even though the Bayes factor will be just ½, the prior probability of H_0 is likely to be small (0.1 or less). This strikes us as condoning the SUTC of the prior probability of H_0. The P-value is reported and stated to provide significant evidence against H_0, with no mention of the fact that this is sensible only because the prior probability of H_0 is very small. Can this really be superior to separately reporting the prior probability of H_0 and the Bayes factor?

As a specific example, let us add two rows to Table 1 of Casella and Berger: here π_0 is the prior probability of H_0, and B is the Bayes factor against H_0.

TABLE 1 (modified)

x	1.645	1.96	2.576	2.807	3.29	3.89
P-value	.10	.05	.01	.005	.001	.0001
$\varepsilon = \varepsilon^*$	.257	.221	.173	.160	.138	.117
$P(\lvert\theta\rvert \le \varepsilon \mid x)$	.079	.043	.011	.006	.002	.0003
π_0	.102	.088	.069	.064	.055	.047
B	1/1.32	1/2.15	1/6.66	1/11.33	1/29	1/164

We would argue that reporting only the P-value or $P(\lvert\theta\rvert \le \varepsilon \mid x)$ is grossly inadequate, hiding the fact that much of the evidence against H_0 is due to π_0 being so small. It may be perfectly reasonable to have π_0 small, but others certainly have a right to know that this was an important component of the conclusion.

Another problem with the argument of Casella and Berger is that, even if it is true that the prior probability of H_0 is typically small, shouldn't it matter whether this prior probability is 0.1 or 0.2 or 0.05 or 0.01? The P-value will only typically correspond to one of these, and it is a bit hard to single out a specific prior probability as always appropriate.

In the same vein, the calculations made in Table 1, Figure 1 and Table 2 are highly arbitrary. By choosing various small interval sizes (we were *not* saying that the ε^* we discussed gave, in any sense, a "typical small interval") and priors $g(\theta)$ one could produce virtually any posterior probabilities whatsoever. Surely Casella and Berger do not mean to imply that it will almost always happen, by chance, that the interval size and g in a precise hypothesis test will be such that the posterior probability and P-value are equal, especially since these will also have to miraculously match up with the sample size in the right way.

It is tempting to go on and on, but in some sense the whole argument is irrelevant. We had focused on the Bayes factor in the paper in the hope of avoiding this issue. The fundamental equation

$$\frac{P(H_0 \mid x)}{P(H_1 \mid x)} = \frac{\pi_0}{(1-\pi_0)} \cdot \frac{1}{B},$$

shows that the posterior odds are affected by the data only through B, which in turn does not involve π_0, the prior probability of H_0. Focusing on what the data has to say, through B, makes concerns as to which π_0 will tend to be appropriate in practice rather irrelevant.

To emphasize the importance of this point, it is helpful to imagine that we had written the paper without ever mentioning π_0 or posterior probabilities, and had considered only the Bayes factor. *All* the arguments and conclusions we draw would remain unchanged, yet the comments of Casella and Berger would now be vacuous. Attention in these arguments should always be focused on the Bayes factor.

Specific Remarks

(i) "Contrary to what Berger and Delampady would have us believe, a great many practitioners should not be testing point nulls, but should be setting up confidence intervals," state Casella and Berger. This is a misrepresentation of what we (Berger and Delampady) would have you believe. A confidence interval for θ, say, will virtually always be necessary to provide knowledge of where θ is, should H_0 be false, and also to judge whether θ is far enough from θ_0 to make a practical difference. Our only point was that a confidence set cannot necessarily be used to reject a believable precise null hypothesis. Thus, in our Rejoinder 3, we observed that a minimal report will often be *both* the Bayes factor against H_0 *and* a confidence set for θ (conditional on H_0 being false); we never implied that the Bayes factor alone would be sufficient for all statistical questions.

(ii) In regards to the third type of precise hypothesis introduced by Casella and Berger, we certainly meant to include this in our type (1). We meant "convenience" to include the type of considerations that they raise, and not just misspecified hypotheses. But we were vague and Casella and Berger are clear; thus, let's grant the fundamental importance of the type (3) hypothesis and proceed. *Point* 1. This type of hypothesis is distinct from type (2), which was the focus of our article, so we could simply say "different situation" and stop. *Point* 2. "as Berger and Delampady admit in Section 5, P-values are reasonable measures of evidence when there is no a priori concentration of belief about H_0" state Casella and Berger. We indeed said it, and we now regret it. The Bayes factor against θ_0 does not, in any way, depend on the prior believability of H_0, and so, if one is trying to determine if θ_0 is compatible with the data or not, the issue of prior believability is irrelevant. We are not arguing that the Bayes factor necessarily answers the questions of interest here; the real questions of interest are probably best answered by approaches such as those discussed in Bayarri's comments. But a P-value is even less likely to answer the questions of interest.

(iii) The claim that "the Bayesian can use the P-value as an approximate posterior probability for large n, regardless of the value of π_0" is not really true. The calculation in Section 2.3, that $P(H_0 \mid \bar{x}_n) \to \alpha$, was based on (a) assuming that ε_0 is known (and nonzero); (b) assuming that $\bar{x}_n = \varepsilon_0 + z_\alpha \sigma/\sqrt{n}$ (so that the P-value remains fixed at α); and (c) letting $n \to \infty$. In practice it is rare to be able to precisely pin down ε_0, and one cannot say, for any given $\bar{x}_n$ and n, whether $P(H_0 \mid \bar{x}_n) \cong \alpha$, without knowing the prior.

REPLY TO KADANE

Our attitudes concerning what is important are often shaped by the problems we encounter in application of statistics. Professor Kadane provides an enjoyable account of his own experiences, and concludes that we should forget about testing altogether. In contrast, we recall seeing a videotape of Sir Harold Jeffreys several years ago, in which he stated that his most important statistical contribution was the development of a Bayesian version of significance testing. Apparently Kadane and Jeffreys have worked on very different problems.

Two more specific responses are: (i) we demonstrated that interval nulls of width $\frac{1}{2}\sigma/\sqrt{n}$ or smaller can effectively be treated as exact point nulls, and such interval nulls may be somewhat more common in Kadane's experience; and (ii) as pointed out by Casella and Berger, there are often particularly interesting values, θ_0, that one might care to determine the evidence (say, Bayes factor) against, even if the values are not specifically believable hypotheses. Of course, we have already admitted that the Bayarri type of calculation will tend to be more relevant for such problems.

ADDITIONAL REFERENCE

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