DISCUSSION

E.T. JAYNES (*Washington University*):

It is always interesting to recall the arguments that Jeffreys used to find priors. The case recounted by Zellner is a typical example where it appears at first glance that we have nothing to go on; yet by thinking more deeply, Jeffreys finds something. He shows an uncanny ability to see intuitively the right thing to do, although the rationalization he offers is sometimes, as Laplace said of Bayes' argument, "fine et très ingénieuse, quoiqu'un peu embarrassée". It was from studying these flashes of intuition in Jeffreys that I become convinced that there must exist a general formal theory of determination of priors by logical analysis of the prior information-and that to develop it is today the top priority research problem of Bayesian theory.

Pragmatically, the actual results of the Jeffreys-Zellner-Siow and Bernardo tests seem quite reasonable; without considerable analysis one could hardly say how or whether we should want them any different. Likewise, there is little to say about the mathematics, since once the premises are accepted, all else seems to follow in a rather straightforward and inevitable way. So let us concentrate on the premises; more specifically, on the technical problems encountered in both works, caused by putting that lump of prior probability on a single point $\lambda = 0$.

1. *The problem*

In most Bayesian calculations the same prior appears in numerator and denominator, and any normalization constant cancels out. Usually, passage to the limit of an "uninformative" improper prior is then uneventful; i.e., our conclusions are very robust with respect to the exact prior range. But in Jeffreys' significance test this robustness is lost, since $K = p(D \mid H_o)/p(D \mid H_1)$ contains in the denominator an uncancelled factor which is essentially the prior density $\pi(\lambda)$ at $\lambda = x$. Then in the limit of an improper prior we have $K \to \infty$ independently of the data D, a result given by Jeffreys (1939, p. 194, Eq. 10), and since rediscovered many times. Note that the difficulty is not due solely to the different dimensionality of the parameter spaces; it would appear in any problem where we think of H_o as specifying a definitive, fixed prior range, but fail to do the same for H_1.

Jeffreys (1961) dealt with this and other problems by using a Cauchy prior $\pi(\lambda \mid \sigma)$ scaled on σ in the significance test, although he would have used a uniform prior $\pi(\lambda) = 1$ in the same model H_1 had he been estimating λ. But then a question of principle rears up. To paraphrase Lindley's rhetorical question: Why should our prior knowledge, or ignorance, of λ depend on the question we are asking about it? Even

more puzzling: why should it depend on another parameter σ, which is itself unknown? One feels the need for a clearer rationalization.

Furthermore, the difficulty was not really removed, but only concealed from view, by Jeffreys' procedure. All his stated conditions on the prior would have been met equally well had he chosen a Cauchy distribution with interquartile span 4σ instead of σ; but then all his K-values would have been quadrupled, leading to indifference at a very different value of the t-statistic [see Eq. (5-13) below]. We do not argue that Jeffreys made a bad choice; quite the contrary. Our point is rather that in his choice there were elements of arbitrariness, arising from a still unresolved question of principle. Pending that resolution, one is not in a position to say much about the "uniqueness" or "objectivity" of the test beyond the admitted virtue of yielding results that seem reasonable.

Bernardo comes up against just the same problem, but deals with it more forthrightly. Finding again that the posterior probability P_o of the null hypothesis H_o increases with the prior variance σ_1 in a disconcerting way, he takes what I should describe as a meat-axe approach to the difficulty, and simply chops away at its prior probability p until $P_o = pk/(pk + 1\text{-}p)$ is reduced to what he considers reasonable (from the Jeffreys-Zellner-Siow standpoint he chops a bit too much, since his P_o tends only to 1/2 on prolonged sampling when H_o is true). This approach has one great virtue: whereas the Jeffreys results tended to be analytically messy, calling for tedious approximations, Bernardo emerges triumphantly (in the limit of large σ_1) with a beautifully neat expression (Eq. (11)) which has also, intuitively, a clear ring of truth to it.

But for this nice result, Bernardo pays a terrible price in unBayesianity. He gets it only by making p vary with the sample size n, calling for another obvious paraphrase of Lindley. This elastic quality of his prior is rationalized by an information-theoretic argument; it is, in a sense, the prior for which one would expect (before seeing the data) to learn the most from the experiment. But is this the property one wants?

If a prior is to incorporate the *prior information* we had about λ before the sample was observed, it cannot depend on the sample. The difficulty is particularly acute if the test is conducted sequentially; must we go back to the beginning and revise our prior as each new data point comes in? Yet after all criticisims I like the general tone of Bernardo's result, and deplore only his method of deriving it.

The common plot of these two scenarios is: we (1) start to apply Bayes' theorem in what seems a straightforward way; (2) discover that the result has an unexpected dependence on the prior; (3) patch things up by tampering with the prior until the expected kind of result emerges. The Jeffreys and Bernardo tamperings are similar in effect, although they offer very different rationalizations for what they do. But in both cases the tampering has a mathematical awkwardness and the rationalization a certain contrived quality, that leads one to ask whether some important point has been missed.

Now, why should that first result have been unexpected? If, according to H_1, we know initially only that λ is in some very wide range $2\sigma_1$, and we then receive data showing that it is actually within $\pm\sigma/\sqrt{n}$ of the value predicted by H_o, -as a physicist would put it, "the data agree with H_o to within experimental error"- that is indeed very strong evidence in favor of H_o. Such data ought to yield a likelihood ratio $K = \sqrt{n}\sigma_1/\sigma$

increasing with σ_1, just as Bernardo finds. This first result is clearly the correct answer to the question Q_1 that was being asked.

If we find that answer disconcerting, it can be only because we had in the back of our minds a different, unenunciated question Q_2. On this view, the tampering is seen as a mutilation of equations originally designed to answer Q_1, so as to force them to answer instead Q_2.

The higher-level question: "Which question should we ask?" does not seem to have been studied explicitly in statistics, but from the way it arises here, one may suspect that the answer is part of the necessary "software" required for proper use of Bayesian theory. That is, just as a computer stands ready to perform any calculation we ask of it, our present theory of Bayesian inference stands ready to answer any question we put to it. In both cases, the machine needs to be programmed to tell it which task to perform. So let us digress with some general remarks on question-choosing.

2. *Logic of Questions*

For many years I have called attention to the work on foundations of probability theory by R.T. Cox (1946,1961) which in my view provides the most fundamental and elegant basis for Bayesian theory. We are familiar with the Aristotelian deductive logic of propositions; two propositions are equivalent if they say the same thing, from a given set of them one can construct new propositions by conjunction, disjunction, etc. The probability theory of Bernouilli and Laplace included Aristotelian logic as a limiting form, but was a mathematical extension to the intermediate region $(0<p<1)$ between proof and disproof where, of necessity, virtually all our actual reasoning takes place. While orthodox doctrine was rejecting this as arbitrary, Cox proved that it is the only consistent extension of logic in which degrees of plausibility are represented by real numbers.

Now we have a new work by Cox (1978) which may prove to be of even more fundamental importance for statistical theory. Felix Klein (1939) suggested that questions, like propositions, might be used as logical elements. Cox shows that in fact there is an exactly parallel logic of questions: two questions are equivalent if they ask the same thing, from a given set of them one can construct new questions by conjunction (ask both), disjunction (ask either), etc. All the "Boolean algebra" of propositions may be taken over into a new symbolic algebra of questions. Every theorem of logic about the "truth value" of propositions has a dual theorem about the "asking value" of questions.

Presumably, then, besides our present Bayesian statistics -a formal theory of optimal inference telling us which propositions are most plausible- there should exist a parallel formal theory of optimal inquiry, telling us which questions are most informative. Cox makes a start in this direction, showing that a given question may be defined in many ways by the set of its possible answers, but the question possesses an entropy independent of its defining set, and the entropies of different questions obey algebraic rules of combination much like those obeyed by the probabilities of propositions.

The importance of such a theory, further developed, for the design of experiments and the choosing of procedures for inference, is clear. For over a century we have

argued over which *ad hoc* statistical procedures ought to be used, not on grounds of any demonstrable properties, but from nothing more than ideological committments to various preconceived positions. There is still a great deal of this in my exchanges with Margaret Maxfield and Oscar Kempthorne in Jaynes (1976), and even a little in the exchange with Dawid, Stone, and Zidek over marginalization in Jaynes (1980). A formal theory of optimal inquiry might resolve differences of opinion in a way that Wald-type decision theory and Shannon-type information theory have not accomplished.

Our present problem involves a special case of this. If, seeing the answer to question Q_1 we are unhappy with it, what alternative question Q_2 did we have, unconsciously, in the back of our minds? Is there a question Q_3 that is the optimal one to ask for the purpose at hand? Since the conjectured formal theory of inquiry is still largely undeveloped, we try to guess some of its eventual features by studying this example.

Note that the issue is not which question is "correct". We are free to ask of the Bayesian formalism any question we please, and it will always give us the best answer it can, based on the information we have put into it. But still, we are in somewhat the position of a lawyer at a courtroom trial. Even when he has on the stand a witness who knows all the facts of the case and is sworn to tell the truth, the information he can actually elicit from this witness still depends on his adroitness in asking the right questions.

If his witness is unfriendly, he will not extract any information at all unless he knows the right questions to force it out, phrasing them as sharp leading questions and demanding unequivocal "yes" or "no" answers. But if a witness is friendly and intelligent, one can get all the information desired more quickly by asking simply, "Please tell us in your own words what you know about the case?" Indeed, this may bring out unexpected new facts for which one could not have formulated any specific question.

Significance tests which specify a sharply defined hypothesis and preassigned significance level, and demand to know whether the hypothesis does or does not pass at that level, therefore in effect treat probability theory as an unfriendly witness and automatically preclude any possibility of getting more information than that one bit demanded.

Suppose we try instead the opposite tactic, and regard Bayesian formalism as a friendly witness, ready and willing to give us all the pertinent information in our problem even information that we had not realized was pertinent if we only allow if the freedom to do so. Instead of demanding the posterior probability of some sharply formulated null hypothesis H_0, suppose we ask of it only, "Please tell us in your own words what you know about λ?" Perhaps by asking a less sharp and restrictive question, we shall elicit more information.

3. *Information from questions*

Evidently, to deal with such problems one ought to be an information theorist, and not only in the narrow sense of One-Who-Uses-Entropy. In the present problem we are concerned not only with the range of possible answers, as measured by the

entropy of a question, but also with the specific kind information that the question can elicit. In the following we use the word "information" in this semantic sense rather than the entropy sense.

All statistical procedures are in the last analysis prescriptions for information processing: what information have we put into our mathematical machine, and what information are we trying to get out of it? In these terms, what is the difference -if any- between significance testing and estimation? Having put certain information (model, prior, and data) into our hopper, we may carry out either, by asking different questions. But the answers to different questions do not necessarily convey different information.

The tests considered by Zellner and Bernardo sought information that can help us decide whether to adopt a new hypothesis H_1 with a value of λ different from its currently supposed value $\lambda = 0$. Presumably, any procedure which yields the same information would be equally acceptable for this purpose, even though current pedagogy might not call it a "significance test".

Now this information criterion establishes an ordering of different procedures, or "tests", rather like the notion of admissibility. If test B (which answers question Q_B) always gives us the same information as test A, and sometimes more, then B may be said to dominate A in the sense of information yield, or question Q_B dominates Q_A in "asking power"; and if B requires no more computation, on what grounds could one ever prefer A?

In my work of 1976 (p. 185 and p. 219), I showed that the original Bayesian sigificance test of Laplace, which asks for the posterior probability P_1 of a one-sided alternative hypothesis, dominates the traditional orthodox t-test and F-test in just this sense. That is, given P_1 we know what the verdict would be, at any significance level, for all three of the corresponding orthodox tests (one equal-tails and two one-sided; but the veredict of any one orthodox test is far from determining P_1. Thanks to Cox, we have now a much broader view of this phenomenon.

Let us call a question *simple* if its answer is a single real number; or in Cox's terminology, if its irreducible defining set is a set of real numbers. For example: "What is the probability that λ, or some function of λ, lies in a certain region R?"

In any problem involving a single parameter λ for which there is a single sufficient statistic u, then given any simple question Q_A about λ, the answer will be, necessarily, some function $a(u)$. Given any two such questions Q_A, Q_B and any fixed prior information, the answers $a(u)$, $b(u)$, being functions of a single variable u, must obey some functional relation $a = f(b)$. If $f(b)$ is single-valued, then the answer to Q_B tells us everything that the answer to Q_A does. As Cox puts it, "An assertion answering a question answers every implicate of that question". If the inverse function $b = f^{-1}(a)$ is not single-valued, then Q_B dominates Q_A.

In the case of a single sufficient statistic, then, any simple question whose answer is a strict monotonic function of u, yields all the information that we can elicit about λ, whatever question we ask; and it dominates any simple question whose answer is not a strict monotonic function of u. But this is just the case discussed by Bernardo; he considers σ known, and consequently $\bar{x}$ is sufficient statistic for λ. Since his odds ratio $K(x)$ is not a strict monotonic function of x, we know at once that Bernardo's test is

dominated by another.

The Jeffreys-Zellner-Siow tests are more subtle in this respect, since σ is unknown, and consequently there are two jointly sufficient statistics $(\bar{x},s)$. Given two simple questions Q_A, Q_B with answers $a(\bar{x},s)$, $b(\bar{x},s)$, the condition that they ask essentially the same thing, leading to a functional relation $a=f(b)$, is that the Jacobian $J = \partial(a,b)/\partial(\bar{x},s)$ should vanish. If $J \neq 0$, then neither questions can dominate the other and no simple question can dominate both. But any two simple questions for which $(\bar{x},s)$ are uniquely recoverable as single-valued functions $\bar{x}(a,b)$, $s(a,b)$ will jointly elicit all the information that any question can yield, and thus their conjunction dominates any simple question.

We may, therefore, conclude the following. Since Jeffreys' test asks a simple question, whose answer is the odds ratio $K(\bar{x},s)$, it can be dominated by a compound question, the conjuction of two simple questions. Indeed, since K depends only on the magnitude of the statistic t, it is clear that Jeffreys' question is dominated by any one simple question whose answer is a strict monotonic function of t.

These properties generalize effortlessly to higher dimensions and arbitrary sets. Whenever sufficient statistics exist, the most searching questions for any statistical procedure, -whatever current pedagogy may call it- are those (simple or compound) from whose answers the sufficient statistics may be recovered; and all such questions elicit just the same information from the data.

As soon as I realized this, it struck me that this is exactly the kind of result that Fisher would have considered intuitively obvious from the start; however, a search of his collected works failed to locate any passage where such an idea is stated. Perhaps others may recall instances where he made similar remarks in private conversation; it is difficult to believe that he was unaware of it.

With these things in mind, let us re-examine the rationale of the Jeffreys-Zellner-Siow and Bernardo tests.

4. *What is our rationale?*

In pondering this -trying to see where we have confused two different questions and what the question Q_2 is- I was struck by the constrast between the reasoning used in the proposed tests and the reasoning that physicists use, in everyday practice, to decide such matters. We cite one case history; recent memory would yield a dozen equally good, which make the same point.

In 1958, Cocconi and Salpeter proposed a new theory H_1 of gravitation, which predicted that the inertial mass of a body is a tensor. That is, instead of Newton's $F=Ma$, one had $F_i = \Sigma M_{ij} a_j$. For terrestrial mechanics the principal axes of this tensor would be determined by the distribution of mass in our galaxy, such that with the x-axis directed toward the galactic center, $M_{xx}/M_{yy} = M_{xx}/M_{zz} = (1+\lambda)$. From the approximately known galactic mass and size, one could estimate (Weisskopf, 1961) a value $\lambda \cong 10^{-8}$.

Such a small effect would not have been noticed before, but when the new hypothesis H_1 was brought forth it became a kind of challenge to experimental physicists: devise an experiment to detect this effect, if it exists, with the greatest possible sensitivity. Fortunately, the newly discovered Mössbauer effect provided a test

with sensitivity far beyond one's wildest dreams. The experimental verdict (Sherwin, *et.al,* 1960) was that λ, if it exists, cannot be greater than $|\lambda| < 10^{-15}$. So we forgot about H_1 and retained our null hypothesis: H_o = Einstein's theory of gravitation, in which $\lambda = 0$.

From this and other case histories in which other conclusions were drawn, we can summarize the procedure of the physicist's significance test as follows: (A) Assume the alternative H_1, which contains a new parameter λ, true as a working hypothesis. (B) On this basis, devise an experiment which can measure λ with the greatest possible precision. (C) Do the experiment. (D) Analyze the data as a pure estimation problem-Bayesian, orthodox, or still more informal, but in any event leading to a final "best" estimate and a statement of the accuracy claimed: $(\lambda)_{est} = \lambda' \pm \delta\lambda$. It is considered good form to claim an accuracy $\delta\lambda$ corresponding to at least two, preferably three, standard deviations. (E) Let λ_o be the correct value according to the null hypothesis H_o (we supposed $\lambda_o = 0$ above, but it is now best to bring it explicitly into view), and define the "statistic" $t \equiv (\lambda' - \lambda_o)/\delta\lambda$. Then there are three possible outcomes:

If $\|t\| < 1$, retain H_o,	STATUS QUO
If $\|t\| >> 1$, accept H_1,	AWARD NOBEL PRIZES
If $1 < \|t\| < 3$, withhold judgment	SEEK BETTER EXPERIMENTS

That is, to within the usual poetic license, the reasoning format in which the progress of physics takes place.

You see why I like the actual results reported here by Zellner and Bernardo, although I find their rationalizations puzzling. They did indeed find, as the criterion for accepting H_1, that the estimated deviation $|\lambda' - \lambda_o|$ should be large compared to the accuracy of the measurement, considered known ($\sigma/\sqrt{n}$) in Bernardo's problem, and estimated from the data in the usual way ($s/\sqrt{n}$) in Zellner's.

It is in the criterion for retaining H_o that we seem to differ; contrast the physicist's rationale with that usually advanced by statisticians, Bayesian or otherwise. When we retain the null hypothesis, our reason is not that it has emerged from the test with a high posterior probability, or even that it has accounted well for the data. H_o is retained for the totally different reason that if the most sensitive available test fails to detect its existence, the new effect $(\lambda - \lambda_o)$ can have no observable consequences. That is, we are still free to adopt the alternative H_1 if we wish to; but then we shall be obliged to use a value of λ so close to the previous λ_o that all our resulting predictive distributions will be indistinguishable from those based on H_o.

In short, our rationale is not probabilistic at all, but simply pragmatic; having nothing to gain in predictive power by switching to the more complicated hypothesis H_1, we emulate Ockham. Note that the force of this argument would be in no way diminished even if H_o had emerged from some significance test with an extremely low posterior probability; we would still have nothing to gain by switching. Our acceptance of H_1 when $|t| >> 1$ does, however, have a probabilistic basis, as we shall see presently.

Today, most physicists have never heard the term "significance test". Nevertheless, the procedure just described derives historically from the original tests devised by Laplace in the 18'th Century, to decide whether observational data indicate

the existence of new systematic effects. Indeed, the need for such tests in astronomy was the reason why the young Pierre Simon developed an interest in probability theory, forty-five years before he became the *Marquis de Laplace*. This problem is therefore the original one, out of which "Bayesian statistics" grew.

As noted also by E.C. Molina (1963) in introducing the photographic reproduction of Bayes' paper, even the result that we call today "Bayes' theorem" was actually given not by Bayes but by Laplace (the only valid reason I have found for calling it "Bayes' theorem" was provided at this meeting; "There's no theorem like Laplace's theorem" does not set well to Irving Berlin's music). Molina also offers some penetrating remarks about Boole's work, showing that those who have quoted Boole in support of their criticisms of Bayes and Laplace may have mistaken Boole's intention.

Now, although Laplace's tests were thoroughly "Bayesian" in the sense just elucidated, they encountered no such difficulty as those found by Jeffreys and Bernardo; he always got clear-cut decisions from uniform priors without tampering. To see how this was managed, let us examine the simplest of all Laplacian significance tests.

As soon as fairly extensive birth records were kept, it was noticed that there were almost always slightly more boys than girls, the ratio for large samples lying usually in the range $1.04 < (n_b/n_g) < 1.06$. Today we should, presumably, reduce this to some hypothesis about a difference in properties of X and Y chromosomes (for example, the smaller Y chromosome, leading to a boy, would be expected to migrate more rapidly). But for Laplace, knowing nothing of such things, the problem was much simpler. Making no reference to any causal mechanism, he took the model of Bernoulli trials with parameter λ = probability of a boy.

His problem was then: given specific data $D = \{n_b, n_g\}$, do these data indicate the existence of some systematic cause favoring boys? Always direct and straightforward in his thinking, for him the proper question to ask of the theory was simply: Q_L = "Conditional on the data, what is the probability that $\lambda > (1-\lambda)$?" With uniform prior, answer was

$$P_L = \frac{(n+1)!}{n_b!\, n_g!} \int_{\lambda_o}^{1} \lambda^{n_b}(1-\lambda)^{n_g} d\lambda$$

with $n = n_b + n_g$, $\lambda_o = 1/2$. In this *Essai Philosophique* Laplace reports many results from this, and in the *Theorie Analytique* (Vol. 2, Chap. 6) he gives the details of his rather tedious methods for numerical evaluation.

Needless to say, Laplace was familiar with the normal approximation to $p(d\lambda \mid D)$, the inverse of the de Moivre-Laplace limit theorem. But Laplace also realized that the normal approximation is valid only within a few standard deviations of the peak, and when the numbers n_b, n_g become very large, it can lead easily to errors of a factor of 10^{100} in $P_L/(1-P_L)$; hence his tedious methods.

Bernardo's example of Mrs. Stewart's telepathic powers, where the null hypothesis value $\lambda_o = 0.2$ is about 24 standard deviations out, is another instance where the normal approximation leads to enormous numerical errors in K (many millions, by my estimate).

But pragmatically, once it is estimated that an odds ratio is about 10^{130}, it hardly matters if the exact value is really only 10^{120}. Once it is clear that the evidence is overwhelmingly in favor of H_1, nobody cares precisely how overwhelming it is. After Laplace's time, physicists lost interest in his accurate but tedious evaluations of P_L; for the criterion that we have overwhelming evidence in favor of a positive effect ($\lambda > \lambda_o$), is just that the overwhelmingly greater part of the mass of the posterior distribution $p(d\lambda \mid D)$ shall lie to the right of λ_o. In the above example, the peak and standard deviation of $p(d\lambda \mid D)$ are $\lambda' = n_h/n$, $\delta\lambda = [\lambda'(1-\lambda')/n]^{1/2}$ and this criterion reduces to the aforementioned $t = (\lambda' - \lambda_o)/\delta\lambda >> 1$, of the modern physicist's significance test-just the same criterion that Jeffreys and Bernardo arrive at in their different ways.

We have noted above that the orthodox t-test and F-test are dominated by Laplace's, and argued that the Jeffreys and Bernardo tests must also be dominated by some other. Let us now compare their specific tests with the ones Laplace would have used in their problems.

5. *Comparisons with Laplace*

In Bernardo's problem we have a normal sampling distribution $p(dx \mid \lambda,\sigma) \sim N(\lambda,\sigma)$ with σ known. Hypothesis H_o specifies $\lambda = \lambda_o$, H_1 a normal prior $\pi(d\lambda \mid H_1) \sim N(\mu_1,\sigma_1)$, leading to a normal posterior distribution $p(d\lambda \mid D,H_1) \sim N(\lambda',\delta\lambda)$ where

$$(\delta\lambda)^{-2} = n\sigma^{-2} + \sigma_1^{-2} \tag{5.1}$$

$$\lambda' = n(\delta\lambda/\sigma)^2 \bar{x} + (\delta\lambda/\sigma_1)^2\mu_1 \tag{5.2}$$

Laplace, asking for the probability of a positive effect, would calculate

$$P_L = p(\lambda > \lambda_o \mid D,H_1) = \Phi(t) \tag{5.3}$$

where $\Phi(t)$ is the cumulative normal distribution, and as always, $t \equiv (\lambda' - \lambda_o)/\delta\lambda$.

Bernardo (Eq. 9) finds for the posterior odds ratio

$$K_B = p(H_o \mid D)/p(H_1 \mid D) = \exp(-R/2) \tag{5.4}$$

where

$$R = \frac{(\bar{x}-\lambda_o)^2}{\sigma^2/n} - \frac{(\bar{x} - \mu_1)^2}{\sigma_1^2 + \sigma^2/n} . \tag{5.5}$$

But by algebraic rearrangement, we find this is equal to

$$R = t^2 - w^2 \tag{5.6}$$

where $w \equiv (\mu_1 - \lambda_o)/\sigma_1$ is independent of the data and drops out if $\mu_1 = \lambda_o$ or if $\sigma_1 \to \infty$. Bernardo would then find for the posterior probability of the null hypothesis

$$P_B = p(H_o|D) = [\exp(t^2/2) + 1]^{-1} \tag{5.7}$$

and comparing with (5.3) we have, as anticipated, a functional relation $P_B = f(P_L)$. To see the form of it, I plotted P_B against P_L and was surprised to find a quite accurate semicircle, almost as good as one could make with a compass. To all the accuracy one could use in a real problem, the functional relation is simply

$$P_B \cong [P_L(1-P_L)]^{1/2}, \qquad 0 \le P_L \le 1 \tag{5.8}$$

The error in (5.8) vanishes at five points ($0 \le P_L \le 1$).

Since $P_B = f(P_L)$ is single-valued while the inverse function is not, we have the result that Laplace's original significance test does, indeed, dominate Bernardo's. As stressed in Jaynes (1976), one-sided tests always dominate two-sided ones; gives P_L we know everything that Bernardo's K or P_B can tell us; and if $|t| >> 1$ we know in addition whether $\lambda > \lambda_o$ or $\lambda < \lambda_o$, which P_B does not give.

Of course, in this case one can determine that extra bit of information from a glance at the data; so the mere fact of domination is hardly a strong selling point. What is important is that Laplace's method achieves this without any elements of arbitrariness or unBayesianity.

In Jeffreys' problem we have the same sampling distribution, with the standard likelihood function $L(\lambda,\sigma) = \sigma^{-n} \exp[-ns^2Q^2(\lambda)/2\sigma^2]$, where

$$Q(\lambda) \equiv [1 + (\lambda-\bar{x})^2/s^2]^{1/2} \tag{5.9}$$

H_0 and H_1 assign common priors $d\sigma/\sigma$, but H_o specifies $\lambda = \lambda_o$, while H_1 assigns the Cauchy prior $p(d\lambda|\sigma,H_1) = \pi(\lambda|\sigma)d\lambda$ with the density

$$\pi(\lambda|\sigma) = \frac{a\sigma}{\pi(a^2\sigma^2 + \lambda^2)} \tag{5.10}$$

scaled on σ (Jeffreys takes $a = 1$, $\lambda_o = 0$, but we define the problem thus to bring out some points noted in Sec. 1). To analyze the import of the data, Jeffreys then calculates the likelihood ratio

$$K_J(x,s) = \frac{p(D|H_o)}{p(D|H_1)} = M^{-1} \int_0^\infty L(\lambda_o,\sigma)d\sigma/\sigma \tag{5.11}$$

while Laplace (if he used the same prior) would calculate instead the probability of a positive effect, given H_1:

$$P_L(x,s) = p(\lambda > \lambda_o|D,H_1) = M^{-1} \int_{\lambda_0}^\infty d\lambda \int_0^\infty d\sigma\, \sigma^{-1}\, \pi(\lambda|\sigma)L(\lambda,\sigma) \tag{5.12}$$

These expressions have a common denominator M, equal to the integral in (5.12) with $\lambda_o = -\infty$.

It is straightforward but lengthy to verify that Jeffreys and Laplace do not ask

exactly the same question; i.e., $J \equiv \partial(K_J,P_L)/\partial(\bar{x},s) \neq 0$. However, they are not very different, as we see on making the same approximation (large n) that Jeffreys makes. Doing the σ-integration in (5.12) approximately, the other integrals may be done exactly, leading to the approximate form

$$K_J \cong [\pi(n\text{-}1)/2]^{1/2}\, a(1+q^2)/Q^n(\lambda_o) \tag{5.13}$$

where $q \equiv (\bar{x}/as)$. This reduces to Jeffreys' result [Zellner's Eq. (2.7) in this volume] when $a = 1$, $\lambda_o = 0$. In the same approximation, Laplace's result is the tail area of a t-distribution with n-2 degrees of freedom:

$$P_L \cong A_n \int_{\lambda_o}^{\infty} d\lambda/Q^{n-1}(\lambda) \tag{5.14}$$

where A_n is a normalization constant. Of course, if Laplace used a uniform prior for λ, he would find instead the usual "Student" result with (n-1) degrees of freedom.

In the limit of an improper prior ($a \to \infty$), K_J diverges as noted in Sec. 1, the original motivation for both the Jeffreys and Bernardo tamperings; but the arbitrary parameter a cancels out entirely from Laplace's leading term, appearing only in higher terms of relative order n^{-1}.

Had we been estimating λ instead, we should find the result $(\lambda)_{est} = \lambda' \pm \delta\lambda$, where $\lambda' = \bar{x}$, $\delta\lambda = s/\sqrt{n}$. But Laplace's result (5.14) is a function only of the statistic $t = (\lambda' - \lambda_o)/\delta\lambda$, and Jeffreys' (5.13) is too for all practical purposes (exactly so if $\lambda_o = 0$, as Jeffreys assumes). Therefore, while considering σ unknown has considerably complicated the mathematics, it does not lead to any real difference in the conclusions. Again, Laplace's test yields the same information as that of Jeffreys, and in addition tells us the sign of $(\lambda\text{-}\lambda_o)$. In all cases -Jeffreys, Bernardo, Laplace, and the modern physicist's test- the condition that the data indicate the existence of a real effect is that $|t| >> 1$.

6. *Where does this leave Q_1?*

In summary it should not, in my view, be considered "wrong" to ask the original question Q_1 = "What is the relative status of H_o and H_1 in the light of the data?" But the correct answer to that question depends crucially on the prior range of λ according to H_1; and so the question appears in the retrospect awkward.

Now the original motivation for asking Q_1, stated very explicitly by Jeffreys, was to provide a probabilistic justification for the process of induction in science, whereby sharply defined laws are accepted as universally valid. But as both Jeffreys and Bernardo note, H_o can never attain a positive posterior probability unless it is given some to start with; hence that "pump-priming" lump of prior probability on a single point $\lambda = 0$. It seems usually assumed that this step is the cause of the difficulty.

However, the question Q_1 is awkward in another, and I think more basic, respect. The experiment cannot distinguish differences in λ smaller than its "resolving power" $\delta\lambda = s/\sqrt{n}$. Yet Q_1 asks for a decision between H_o and H_1 even when $|\lambda\text{-}\lambda_o| < \delta\lambda$. On the other hand, the experiment is easily capable of telling us whether λ is probably greater

or less than λ_o (Laplace's question), but Q_1 does not ask this. In short, Q_1 asks for something which the experiment is fundamentally incapable of giving; and fails to ask for something that the experiment *can* give.

[Incidentally, a "reference prior" based on the Fisher information $i(\lambda)$ is basically a description of this resolving power $\delta\lambda$ of the experiment. That is, the reference prior could be defined equally well as the one which assigns equal probabilities to the "equally distinguishable" subregions of the parameter space, of size $\delta\lambda$. This property is quite distinct from that of being "uninformative", although they happen to coincide in the case of single location and scale parameters].

But what we noted in Sec. 4 above suggests a different view of this. Why does induction need a probabilistic justification if it has already a more compelling pragmatic one? It is for the departures from the previous line of induction (i.e., switching to H_1) that we need -and Laplace gave- a probabilistic justification. Bernardo seems to have sensed this also, in being content with the fact that his $p(H_o \mid D)$ tends only to 1/2 when H_o is true. Once we see that maintenance of the *status quo* requires no probabilistic justification, the original reason for asking Q_1 disappears.

7. *Conclusion*

What both the Jeffreys and Bernardo tamperings achieved is that they managed to extricate themselves from an awkward start and, in the end, succeeded in extracting the same information from the data (but for the sign of λ-λ_o) that Laplace's question Q_L = "What is the probability that there is a real, positive effect?" elicited much more easily. What, then, was that elusive question Q_2? It was not identical with Q_L, and perhaps does not need to be stated explicitly at all; but in Cox's terminology we may take Q_2 as *any implicate of Laplace's question whose answer is a strict monotonic function of* $|t|$.

We have seen how the answers to seemingly very different questions may in fact convey the same information. Laplace's original test elicits all the information that can be read off from Jeffreys' $K_J(x,s)$ or Bernardo's $K_B(x)$. And for all purposes that are useful in real problems, Laplace's P_L may in turn be replaced by the λ' and $\delta\lambda$ of a pure estimation problem. Because of this, I suggest that the distinction between significance testing and estimation is artificial and of doubtful value in statistics-indeed, negative value if it leads to needless duplication of effort in the belief that one is solving two different problems.

D.J. SPIEGELHALTER (*University of Nottingham*):

The papers by Professors Zellner and Siow and Bernardo both suggest reference or 'non-informative' priors for use in Bayes factors, but they produce fundamentally different results. I shall begin by comparing these results, and then discuss the individual merits of the two proposals.

Consider the simple case $x \sim N(\mu, \sigma^2/n)$, $H_o: \mu = \mu_o$, $H_1: \mu \neq \mu_o$. Let $\gamma_x = \sqrt{n}(x-\mu_o)/\sigma$ in the notation of Bernardo, who suggests a Bayes factor (17) in favour of the null of $\exp\{-(\gamma_x^2-1)/2\}$, which has behavior similar to that of a significance test. In the case of unknown variance, Zellner proposes the Jeffreys form (2.7) which for n fairly large is approximately equal to $(\pi n/2)^{1/2} \exp(-\gamma_x^2/2)$. For large n, values of γ_x which would lead Bernardo to just reject H_o, would suggest accepting H_o to Zellner.

This is the Lindley paradox, and investigators of Bayes factors have been divided in their support for this phenomenon. Pro-paradox are Zellner (1971), Lindley (1961), Jeffreys (1961), Dickey (e.g. 1971) and Schwarz (1978), while anti-paradox are Akaike (1978), Atkinson (1978), Box and Kanemasu (1973) for 'post data' Bayes factors, and presumably we should include all users of significance tests. Professor Bernardo suggests that a significance test procedure is appropriate in checking a scientific theory. I would be grateful to both authors for some comments on the appropriate practical situations for these two approaches.

The paradox will cause Professor Zellner's Bayes factor wrongly to accept H_o, if the likelihood is concentrated around the true parameter value lying $0(n^{-1/2})$ from H_o. A Bayesian with a true prior under H_1 would, however, consider this event *a priori* extremely unlikely to occur for large n. Moreover, even if this erroneous choice of H_o did occur, for predictive purposes at least, the error is irrelevant since the true model is only a negligible distance from the null. These arguments for the practical use of 'pro-paradox'' Bayes factors are formalised in Smith and Spiegelhalter (1980).

It remains to examine whether the proposals of Zellner and Bernardo are appropriate choices of non-informative prior, within their respective schools of thought on Bayes factors.

Professor Zellner's paper

It has been said at this meeting that 'everything is in Jeffreys'. Perhaps this is an exaggeration, but this paper gives the impression that this work *would* have been in Jeffreys, if only Jeffreys had got round to extending his work to linear models. I trust the authors will take this comment as a compliment of their work, as it is intended.

I have, however, some reservations about the presence of the $\mathbf{X}^T\mathbf{X}$ matrix in the prior specification (3.7b). Changing a prior according to the sampling design would seem somewhat strange. Consider the example of one-way analysis of variance, in which there are I groups with size $n_1, \ldots n_I$, and let $N=\Sigma n_I$. The null hypothesis is of equal group means $H_o: \mu_1 = \ldots = \mu_I = \mu$ against a general alternative. Then (3.12) provides the Bayes factor.

$$K_{01} = \frac{\sqrt{\pi}}{\Gamma(I/2)} \left(\frac{N-I}{2}\right)^{I-1/2} (1-R^2)^{-(N-I-1)/2}$$

Consider a prior that does not depend on the sampling design. Considerations of invariance suggest $p(\mu \mid \sigma, H_o) \propto \sigma^{-1}, p(\mu_i \mid \sigma, H_1) \propto \sigma^{-1}$ and $p(\sigma) \propto \sigma^{-1}$ as non-informative priors, which lead to a Bayes factor

$$B_{01} = C_I(\Pi n_i/N)^{1/2}(1-R^2)^{-N/2}$$

where C_I is some constant of proportionality to be specified. We may adopt the 'device of imaginary results' (Good, 1950) to suggest a plausible value for C_I. Say we observe $R^2=0$ (equal group sample means), then we would presumably expect $B_{01} \geq 1$ which implies $C_I^2 \geq N/\Pi n_i$. A lower bound is given when $n_1=2, n_i=1$, $i=2,\ldots I$, leading to $C_I^2 \geq (I+1)/2$. Assuming this lower bound for illustrative purposes provides a Bayes

factor

$$B_{01} = [(I+1)\Pi n_i/2N]^{1/2}(1-R^2)^{-N/2}$$

To compare the behavior of K_{01} and B_{01}, let $I = 5$, $n_1 = n, n_2 = \ldots = n_5 = m$. Then

$$K_{01} = 1/3\,(N-5)^2(1-R^2)^{-(N-6)/2}$$

$$B_{01} = m^2 3^{1/2}\{n/(n+4m)\}^{1/2}(1-R^2)^{-N/2}$$

If the design is unbalanced, n being large compared with m, then K_{01} will favour H_o much more than the Bayes factor based on a prior that does not depend on the sampling design. This dependence would appear to be quite important and, as previously mentioned, rather alien to the usual methods of prior specification.

Professor Bernardo's paper

I should first congratulate Professor Bernardo for an ingenious extension of his theory of reference priors to the area of Bayes factors. However, I find the definition (6) of missing information a little forced. If θ has a mixed prior, denoted $p'(\theta)$, should we not seek to maximise $I^{\theta}\{\epsilon, p'(\theta)\}$ with respect to p?

By changing p with n, the author wishes to avoid the situation described by equation (13), in which one accepts H_o as the spread of the prior under the alternative increases. If this prior actually expressed one's beliefs, this behaviour seems quite reasonable. So the objection arises from an inappropriate use of a locally uniform prior, whose ordinate at the likelihood is allowed to go to zero. The problem becomes that of choosing an appropriate ordinate for a locally uniform distribution.

Professor Bernardo's wish to avoid the Lindley paradox would seem appropriate in two contexts at least; when there was a large loss on false rejection of the null, even though the alternative is very close, or when we have strong belief *a priori* in alternatives close to the null. If the latter is true, then this should be modelled in our prior. It can be shown (Smith and Spiegelhalter, 1980) that if the prior shrinks around the null at the same rate as the likelihood concentrates, then one obtains a Bayes factor B_{01} which approximately satisfied

$$-2\log_e B_{01} = \lambda-(3/2)(p_1-p_o)$$

where p_i is the number of parameters in H_i, λ is the standard likelihood ratio statistic, and $\lambda \sim \chi^2_{p_1-p_o}$ under H_o. The multiplier 3/2 compares with the use of 2 by Akaike (1978), 1 by Box and Kanemasu (1973), and $\log_e(n-p_1)$ by Professor Zellner in expression (3.24).

The example discussed by the author is equivalent to using a multiplier of 1. I am not sure whether the information theoretic argument is to be extended to the general linear model. If so, one should note that the use of a multiplier 1 may lead to a rather strong preference for complex models, since in this case $E[-2\log_e B_{01}] = 0$ and so the probability that the Bayes factor prefers H_1, given H_o is true, is approximately .5

whatever the complexity of the alternative. I suggest a slightly larger multiplier is more appropriate.

H. AKAIKE (*Institute of Statistical Mathematics, Tokyo*):

We are often told that the Bayesian approach is developed for each particular set of data. This means that the sample size is always equal to 1. I see n's, the sample sizes, in both Professor Zellner's and Professor Bernardo's papers. Any aspects within these papers which essentially depend on n may not then be particularly Bayesian.

In the example of Section 1 of Professor Bernardo's paper, we thus assume $n = 1$. This reduces the problem to the choice of $p = p(H_0)$ in relation to the size of σ_1^2, the variance of the prior distribution of the mean μ. For simplicity we assume $\mu_1 = \mu_0 = 0$ and get $p(x \mid H_0) = N(x \mid 0, \sigma^2)$ and $p(x \mid H_1) = N(x \mid 0, \sigma^2 + \sigma_1^2)$. To keep the predictive distribution $p(x) = pN(x \mid 0, \sigma^2) + (1-p)N(x \mid 0, \sigma^2 + \sigma_1^2)$ impartial to both $p(x \mid H_0)$ and $p(x \mid H_1)$ in terms of entropy, we have to assume $\int p(x \mid H_0) \log(p(x)/p(x \mid H_0))\,dx = \int p(x \mid H_1) \log(p(x)/p(x \mid H_1))\,dx$. When $\sigma_1^2 \to \infty$ this will hold only with $p = 0.5$. For this choice of $p = 0.5$, the critical value of x where the posterior probability of H_0 attains 0.5 is almost equal to 2σ for $\sigma_1 = 8\sigma$ and increases to 3σ for $\sigma_1 = 100\sigma$. This seems to suggest that ordinary choice of significance level such as 5% or 1% is fairly reasonable. The fixed choice of the level may further be questioned. But it is now obvious that the ratio σ_1/σ controls the choice.

A.P. DEMPSTER (*Harvard University*):

Both papers are concerned with Bayesian tests of significance. A standard parametric specification depending on parameters (θ, ϕ) is assumed, and the null hypothesis is that θ takes a prespecified "sharp null" value θ_0, while ϕ is unconstrained. Both papers start from the "paradox" of Lindley (1957) who shows that Bayesian testing and tail area testing produce very different judgments when a diffuse prior distribution is assigned to θ given the alternative hypothesis. As sample sizes increase, the diffuse prior implies that the data tend to add much more credence to the null hypothesis relative to standard tail area tests. Both papers develop alternatives to diffuse priors which bring the Bayesian results into relatively close conformity with tail area results. The papers are worthy contributions to theoretical statistics, but in my view irrelevant to statistical practice.

What is probability? Probability does not fall into distinct categories such as subjective, logical, and physical. Any probability model worth using to assess real world uncertainty must command belief, must result from a chain of reasoning, and must not be in clear conflict with known empirical facts. Bernardo appeals to an information-theoretic principle to derive a prior distribution, while Zellner and Siow appeal to plausible postulates originating with Jeffreys. The reasoning behind these derivations is interesting, but there is no way I can commit belief to the resulting prior distributions, since my prior would then depend on the accident of sample size. Also, there is an empirical "how does it work" component to each paper consisting of comparisons with tail area results, and suggesting that the disparity between the Bayesian techniques and standard non-Bayesian practice is rather mild. But, since tail

area tests are not supposed to be Bayesian, the mildness of the disparity is a logical curiosity rather than evidence that the Bayesian models are credible.

What are significance tests for? The procedures called Bayesian significance testing and tail-area significance testing answer logically different questions, so that the use of the term significance testing for both creates semantic confusion reather than substantive controvery.

In connection with his parable of King Hiero's crown, Savage (1962, pp, 29-33) clearly illustrated the need for Bayesian procedures which provide rational choices between sharp null hypothesis and higher dimensional alternatives. I agree with the Bayesian position which says that the advocates of Neyman-Pearson testing theory are in error when they seek to apply their theory to operational decision-making, as in Pearson (1962). The Neyman-Pearson theory makes probabilistic sense only as a theory *about* tail area tests, and is at that an inadequate theory because if fails to come to grips with the mysteries of conditional testing.

The positive aspect of tail-area tests is that they address real questions which come up in the process of developing a formal model to be used either for purposes of scientific insight or operational decision-making. Specifically, they provide one way to ask whether a nominated model appears to conform to the outside world of fact. Tail-area tests ought to be indispensable to Bayesian statisticians wishing to avoid criticism of their models from two directions. Tail area tests which reject can provide signals that modellers, including Bayesians who have already had their prior model elicited, should go back to the drawing board, because the test shows that the data are trying to say something about phenomena not yet captured in the elicited model. In this case, introspection is not enough, an a further look at the real world may be advisable. Tail area tests which accept can serve to point to possible eventualities whose prior probabilities are influenced only minimally by the data, while these same probabilities may exert a serious influence on later Bayesian conclusions or decisions. In this case, introspection may be all there is, and should be given extra effort. For example, I may not have enough data to detect a significant relation between a chemical agent and a human cancer, but once having raised the question I will not lightly brush off the need to put numbers on prior probabilities of small effects.

In summary, Bayesian statisticians who reject tail area testing are correct when they attack its misuse for decision-making, but are in danger of missing the benefits of correct use in their zeal for things Bayesian. "Significance testing" should be excised from Bayesians have enough good things to do without invading neighboring territory.

What does it mean to "test against an alternative hypothesis"? George Barnard argued at the conference, and I supported him, that significance tests can be valid and important when only the null hypothesis is formulated, as in the Daniel Bernouilli example. Fisher rejected the Neyman-Pearson theory which stressed alternative hypotheses because the theory was couched in terms of long run frequencies, whereas in his mind, as in Daniel Bernouilli's, the purpose of significance testing was to interpret a particular data set. Fisher did not use the formal term alternative hypotheses, but he could scarcely have rejected the concept since the very word "null" suggests that significance testing is a backward way to get at alternative hypotheses.

When a significance test gets to be repeatedly used, appears in "how to do it"

books, and becomes distorted by the term "procedure", then there generally is a reasonably well defined set of alternative hypotheses which are substituted for the null hypothesis when the test produces a significant outcome. It is then sensible, I believe, to use the term "testing a null hypothesis against an alternative hypothesis". I believe also that tail area testing is a clumsy mechanism for the purpose. But I have rejected Bayesian "significance testing" as the answer, so what is left?

The obvious answer in the case of simple null and simple alternative hypotheses is to look ar the likelihood ratio in favor of the alternative. If the ratio is 99 to 1 then the null hypothesis can be "rejected" with similar logic to rejection based on a tail area of .01. When the hypotheses are not simple, my suggestion (1973) is to use the posterior distribution of the likelihood ratio, i.e., the posterior distribution which my Bayesian self would use if I adopted the alternative hypothesis, and to reject the null hypothesis if I am reasonably sure, say 60% sure, that the likelihood ratio is at least 99 to 1. This approach produces judgments similar to tail area tests, and so produces practical answers in the same general range as those of Bernardo and of Zellner and Siow.

These papers resolve the Lindley paradox by producing Bayesian procedures where the paradox largely goes away. I prefer to say there never was a paradox largely because the procedures Lindley contrasts were not comparable in the first place. My work (1973) exhibits alternatives to tail area testing which area genuine significance tests, but are likelihood based. They do not require the contrived priors of Bernardo or Zellner and Siow, but do have a Bayesian element which is relatively insensitive to the choice of prior.

J.M. DICKEY (*University College Wales Aberystwyth*):

Professor Zellner in his paper seems to remain true to Jeffreys' conception when extending Jeffreys' Bayes factors to the general linear model. I should like to point out some disagreeable aspects of the method in Jeffreys' simple context, which extend to the general context. Denote the unknown mean and variance for a simple normal sample, $y_1,\ldots,y_n$, by μ and σ^2. One desires to compare the two models,

$$\mathrm{H}: \mu = 0, \qquad \text{versus } H^c: \mu \neq 0$$

As usual, familiar magic words like "knowing little" are used to introduce a particular prior distribution as being worth one's special attention. The idea seems to be to produce an automatic procedure which will be universally accepted. Under H^c, the joint density proposed is

$$p(\mu,\sigma \mid H^c) \doteq \{f(\mu/\sigma)/\sigma\}\{K/\sigma\} \tag{1}$$

where

$$f(\mu) = \{\pi(1+\mu^2)\}^{-1}$$

(I have introduced a multiplicative constant K here and written an approximate

equality, relative to the likelihood function, in the sense of Savage's "precise measurement").

I assume that the first bracketed factor in (1) represents the conditional prior information concerning μ given σ,

$$p(\mu\,|\,\sigma,H^c) \doteq \{\pi(1+\mu^2/\sigma^2)\}^{-1}/\sigma \tag{2}$$

(One could argue that my assumption is unwarranted. But an alternative factorization would need to be given, rather than mere magic words). Thus, the second factor would be the marginal prior density for σ under H^c.

$$p(\sigma\,|\,H^c) \doteq K/\sigma \tag{3}$$

My first complaint is that the integrable conditional density (2) is *very special.* I have heard it said that the choice of scale $1.\sigma$ is made "for convenience". But why not 100 σ "for convenience", or $(.001)\sigma$, or $(11{,}682.49)\ \sigma$? Clearly, the choice should depend on the actual opinion in each application. Should one act against one's opinions and, instead, report a Bayes factor that represents no person's coherent change of opinion?

One may find it difficult thus to specify ones conditional opinion concerning the location conditional on the unknown scale. But what about the *marginal* opinion concerning μ under H^c? Working directly from the joint density (1), we obtain

$$\begin{aligned} p(\mu\,|\,H^c) &= \textstyle\int_0^\infty p(\mu,\sigma\,|\,H^c)d\sigma \\ &\doteq K\textstyle\int_0^\infty f(\mu/\sigma)/\sigma^2\,d\sigma \\ &= K/\,|\mu|\,. \end{aligned} \tag{4}$$

Again, for my second complaint, this is a very special form and may fail to approximate well ones actual prior opinion concerning μ undr H^c, even locally relative to the likelihood function, even with the constant K open to choice.

Under the hypothesis H, the corresponding prior density which was proposed for σ is

$$p(\sigma\,|\,H) \doteq k/\sigma \tag{5}$$

This constrasts with the conditional density obtained from (1) and (4),

$$\begin{aligned} p(\sigma\,|\,\mu,H^c) &= p(\mu,\sigma\,|\,H^c)\,/\,p(\mu\,|\,H^c) \\ &\doteq |\mu|\,f(\mu/\sigma)/\sigma^2, \end{aligned} \tag{6}$$

which has the asymptotic form near H,

$$\lim_{\mu\to 0} p(\sigma\,|\,\mu,H^c) \propto \sigma^{-2} \tag{7}$$

Note, however, that for the new variable $\eta = \mu/\sigma$, η and σ are prior independent

under H^c according to (2), and hence for any value of η,

$$p(\sigma \mid \eta, H^c) \doteq K/\sigma \quad . \tag{8}$$

In particular, (5) and (8) agree for $\eta = 0$, thereby satisfying Savage's condition continuity. (See my discussion to the paper by Professor Smith in these Proceedings). Note that any other point hypothesis, $\mu = \mu_o$, could not be reexpressed in terms of a point hypothesis on η, since $\mu = \mu_o$ means $\eta = \mu_o/\sigma$.

In my paper in press, Dickey (1978), convenient Bayes factors are provided for the normal linear model together with operational methods for use in cases where the likelihood function is more informative than the prior densities. I also treat intersecting hypotheses, as well as nested and unrelated hypotheses.

S. GEISSER(*University of Minnesota*):

In most statistical problems in which one is dealing with a linear regression, the regression arises not from some "true" physical process but largely from a combination of convenience and an adequate fit of the data in hand. The reasons are twofold , first the so-called "true" process governing the data is often very complex and unknown. Secondly, the interest in the data emanates from a need to predict new values rather than to select a "true" physical model. With this view in mind, W. Eddy and I (1979) devised a selection scheme (useful for a variety of situations including linear regression) which is geared to prediction and derives from a Bayes-Non-Bayes methodological compromise. One of its properties, which superficially appears to be unfavorable, is that asymptotically with non-zero probability it can choose a "wrong" higher dimensional model as opposed to a "true" lower dimensional model. However, it turns out that it is approximately equivalent to a Bayesian procedure with penalties (costs or prior weights) that depend on the sample size and the kind of selection error incurred. What this implies is that even if one chooses the higher dimensional model when the lower one is "true", asymptotically there is no loss incurred for predictive purposes. I believe that such procedures are more useful for most problems that occur in statistics than those that are geared only to selecting the true model, because of primary interest in prediction and the fact that our net hasn't really been cast over the "true" alternative.

I.J. GOOD (*Virginia Polytechnic and State University*):

In accordance with a theorem of Abraham Wald, a minimax procedure corresponds to a Bayesian procedure with the 'least favorable" prior. I pointed out in Good (1969) that if expected weight of evidence is taken as the utility (or quasi-utility) measure, then Wald's theorem leads to the Jeffreys invariant prior. (I believe this is equivalent to what Dr. Bernardo describes in terms of maximizing the missing information). It gives an explanation of why the reference prior is invariant with respect to mere changes of notation, and also explains why it cannot be entirely satisfactory: because minimax methods never are entirely satisfactory except possibly against an intelligent opponent. Nature is neither intelligent, nor an opponent, although life is a losing game.

Regarding "Good's paradox", see my contribution to the discussion of Dr. Zellner's paper.

Dr. Bernardo's Table 1, relating tail-area probabilities to Bayes factors, is remarkably consistent with my rough-and-ready rule that a Bayes factor usually lies between $1/(30P)$ and $3/(10P)$ (Good, 1957, p. 863). But, in various applications this formula can be improved; for example, Good and Crook (1974, p. 715), where $N^{1/2}$ comes into the formula.

D.V. LINDLEY (*University College London*):

These two papers bother me. They are extremely thoughtful papers, rich with ideas, yet they fail to adhere to de Finetti's aphorism, "Think about things". If we have a practical problem of data analysis, the quantities have a physical meaning and the scientist knows something about them. He should therefore be encouraged to think about them, or the parameters, and not adopt probability distributions that merely conform to some patterns of ignorance or some formal model. What does he know about θ? Is it really Cauchy? I do not wish to denigrate these papers, for they both help us enormously to understand the way probabilities behave, and are particularly well-written. But, as this conference comes to an end, it does appear to me that we have discussed technicalities too much and that we should balance this necessary activity with some thinking about the real world, not Greek letters.

A. O'HAGAN (*University of Warwick*):

Would Professor Bernardo please explain why he chooses the particular limiting process he used in section 2 to obtain equation (11)? If we simply let $\sigma_1^2 \to \infty$ in (9), holding all other quantities fixed, we will obtain posterior odds

$$\frac{\pi(H_1|D)}{\pi(H_0|D)} \to \exp(1/2\,\gamma_x^2)$$

Equation (11) is a consequence of holding γ_1 fixed, so that μ_1 increases with σ_1. In the next section he uses yet another limiting process to reach equation (17). All three limiting processes end with a uniform prior on $(-\infty,\infty)$. All three posterior odds expressions have the same qualitative large-sample behaviour that Professor Bernardo likes. Yet they will give numerically quite different posterior inferences in practice. How are we to choose between them?

A. ZELLNER (*University of Chicago*):

At the 1976 Fontainebleau Conference on Bayesian Methods, I pointed out that Bernardo's procedure for generating prior distributions makes the form of the prior dependent on the likelihood function's form, that is on the design of the experiment. This point, apparently unrecognized by Bernardo, was particularly disturbing to Bernardo and Lindley. In Lindley's discussion of Harold Jeffreys's presentation at the Econometric Society's World Congress meeting in 1970, he termed such a dependence to be incoherent. Jeffreys's, Box and Tiao's and my procedures for generating priors

also involve a dependence of a prior's form on the form of the likelihood about which I wrote, Zellner (1977, p. 231) "Since the purpose of a MDIP [maximal data information prior] is to allow the information provided by an experiment to be featured [in the posterior distribution], it seems natural that this form of a MDIP *pdf* that accomplishes this objective be dependent on the design of an experiment". It would be interesting to learn about Bernardo's and Lindley's current position on this issue.

While I do not enjoy raising disturbing points, it should be pointed out that Bernardo's odds ratio in equation (17), $\pi(H_1|D)/\pi(H_0|D) = \exp\{1/2(\gamma_x^2-1)\}$, where $\gamma_x = \sqrt{n}(x-\mu_0)/\sigma$ has a fixed (independent of n) lower bound of $e^{-1/2} = 0.606$. This appears unsatisfactory and is not a characteristic of, for example, Jeffreys's posterior odds ratio for the normal mean problem.

REPLY TO THE DISCUSSION

A. ZELLNER *(University of Chicago):*

One main objective of Jeffreys's and our work is to provide a coherent framework within which it is possible to rationalize and criticize empirical practice in comparing and choosing between or among hypotheses, for example in the normal mean case, $\lambda = 0$ and $\lambda \neq 0$. In this case Jeffreys (1979) states that "...astronomers had a rough rule that discrepancies up to $\pm 2\sigma$ were likely to disappear with more information, and those beyond χ^2. I was glad to find that these [results] were usually about what my significance tests gave. At least they showed that the rough rule corresponded fairly well to a connected theory". Also, see Jeffreys (1967, p. 273) for another statement of this rough rule, a form of which Jaynes cites approvingly in his comments. Producing a "connected theory" to rationalize sensible rules and to criticize absurd rules for significance testing is one of Jeffreys's and our main objectives which we deem important and intimately related to "real world" significance testing problems, a point which Lindley fails to appreciate in his comments. That significance testing procedures (and other statistical procedures) in physics, astronomy, economics and other sciences are in need of improvement is apparent to many statisticians.

As regards sharp null hypotheses, for which Dempster and Savage, among others see a need and significance tests, Jeffreys (1963) writes, "Every quantitative law in physics implies a series of significance tests that have rejected numerous possible modifications of the law" (p. 409). Similarly in biology, economics and other sciences, significance testing involving sharp null hypotheses plays an important role. Thus, Good's suggestion to "roll together significance testing and estimation into a single process" is misguided in our opinion and contradicts Jeffreys's, Dempster's, Savage's and other's stated "need for Bayesian procedures which provide rational choices between sharp null hyyotheses and higher dimensional alternatives," as Dempster puts it in his comments.

On Jeffreys's and our use of particular Cauchy priors upon which most of our discussants have commented, some of them have apparently missed the point that one of the reasons for their use is that posterior odds ratios based on them rationalize the rough rules used by physicists, astronomers and others in testing. They can represent

prior views in a number of cases and serve as a useful reference prior in others. As Jaynes notes, their use has "the admitted virtue of yielding results that seem reasonable". Thus, in response to Akaike's thoughtful comment, their use leads to Bayesian results which are applicable to many sets of data --a general objective of theorizing in many areas including statistics. Further, as Jeffreys (1967, p. 272) and we stated on the first page of our paper, more informative and/or different priors can, and should be employed if the particular Cauchy priors are deemed inadequate to represent the available prior information. However, we believe that the Cauchy priors which we employed will be found useful in many applications and do serve as a basis to rationalize and criticize much current practice. For example, in our framework p-values are given an interpretation and the implications of a choice of usual critical values for a test statistic can be appraised. In addition, Jeffreys (1967, p. 275) points out that the value of the invariant (divergence) measure,

$$J = \int \log \frac{dP'}{dP} \, d(P'-P)$$

for the normal mean problem where P refers to the normal distribution with $\lambda = 0$ and $0 < \sigma < \infty$ and P' to the normal distribution with $\lambda \neq 0$ and $0 < \sigma < \infty$ is $J = \lambda^2/\sigma^2$. He notes that taking a uniform prior on $\theta = \arctan(\lambda/\sigma)$, $-\pi/2 < \theta < \pi/2$ yields exactly the particular Cauchy prior for λ/σ which he employs in the normal mean problem.

We now turn to Jaynes's comments. First, we find no "technical problems" caused by putting a "lump of prior probability on a single point $\lambda = 0$". Second, on the question, "Why should our prior knowledge or ignorance, of λ depend on the question we are asking about it?", Jaynes does not recognize that often when a hypothesis $\lambda = 0$ has been suggested, the value 0 is viewed differently from other possible values. Call this prior information I_0. If the value $\lambda = 0$ is not viewed differently from other values, call this prior information I_1. Then for these frequently encountered circumstances there is good reason for the prior distributions $p_0(\lambda \mid I_0)$ and $p_1(\lambda \mid I_1)$ to be different, a fact appreciated by many including Lindley (1965, p. 58 ff.) in his work on testing procedures when prior information is of type I_1.

With respect to the new work of Cox on the logic of questions, we have some doubts about the adequacy of the entropy concept to judge the value of questions. Be that as it may, in our recent work, Zellner and Siow (1979) on the normal mean problem with σ's value unknown, we consider three hypotheses, H_1: $\lambda = 0$. H_2: $\lambda > 0$ and H_3: $\lambda < 0$, with prior probabilities $\pi_1 = 1/2$, $\pi_2 = \pi_3 = 1/4$ and Cauchy priors such as used in our past work, defined over half line $\lambda > 0$ for H_2 and $\lambda < 0$ for H_3. The approximate posterior odds ratios are:

$K_{12} \doteq g(t,\nu)/F(t)$, $K_{13} \doteq g(t,\nu)/F(-t)$ and $K_{23} \doteq F(t)/F(-t)$ where $t = n^{1/2}\, y/s$, $g(t,\nu) = (\pi\nu/2)^{1/2}/(1+t^2/\nu)^{\nu-1/2}$, which is Jeffreys's odds ratio given in (2.7) of our paper under discussion and $F(\cdot)$ is the cumulative normal distribution function. It is then the case that the posterior odds ratio for $\lambda = 0$ and $\lambda \neq 0$ (the union of H_2 and H_3)

is just $g(t,\nu)$, Jeffreys's posterior odds ratio. Thus, as Jaynes ingeniously suggested, consideration of two questions $\lambda = 0$ vs. $\lambda > 0$ and $\lambda = 0$ vs. $\lambda < 0$ will yield Jeffreys's result under the *special* prior probabilities given above. However, the practical differences are negligible in this case. Also, the expression for K_{23} above is very close to the result yielded by what Jaynes calls the Laplacian approach. Therefore, when we apply Jeffreys's approach to the three hypotheses, it produces the Laplacian result K_{23}, *as well as* K_{12}, K_{13} and posterior distributions for parameters under all three hypotheses. However, for one-sided alternatives in practice, it is often unreasonable to assume that $\pi_2 = \pi_3$. Our recent work indicates that taking $\pi_1 = .5$, $\pi_2 = .4$ and $\pi_3 = .1$ yields results close to non-Bayesian "one-tailed" testing results in terms of indifference values of t for this normal mean problem when the sample size is about 20.

On predictive distributions and testing, which Jaynes mentions, it is well known that the posterior odds ratio with prior odds ratio equal to one is equal to a ratio of predictive densities and thus a posterior odds ratio of about one indicates close agreement of the predictive densities under the two hypotheses. Also, Jaynes's consideration of values of $|t|$ in appraising hypotheses fails to take adequate account of the role of sample size in evaluating hypotheses. Further, when $|t| << 1$, the important result is that the simpler model (e.g. $\lambda = 0$) can be retained. This is important since it is well known that use of models with redundant or unneeded parameters results in inflation of the mean square error of prediction. Thus, in disagreement with Jaynes, there is obviously something valuable to gain in switching to a simpler model when warranted.

Jaynes remarks that Laplace got "clear-cut decisions from uniform priors". Jeffreys's (1967, p. 128 ff.) discussion of Broad's application of Laplace's rule of succession is relevant. In this case, a uniform prior led to unsatisfactory results in a very basic problem. Jeffreys (1967) comments that, "We really had the simplest possible significance test in our modification of Laplace's theory of sampling, where we found that to get results in accordance with ordinary thought we had to suppose an extra fraction of the initial probability, independent of the size of the class, to be concentrated in the extreme values". (p. 247). See also Geisser's (1978) discussion of this problem. Thus for Jeffreys to get sensible results, it was necessary to use "lumps of probability" on extreme values. Finally, it is surprising to us that Jaynes and Good are apparently in disagreement with Jeffreys and many other scientists and statisticians on the need to distinguish significance testing and estimation.

Spiegelhalter aligns researchers with respect to Lindley's "paradox". This appears to us to be a mistake since there is little paradoxical about Lindley's results. As the sample size increases, good sampling theorists will adjust their significance level in an obvious direction, as pointed out in Zellner (1971, p. 304, fn.) and hence no paradox. Good Bayesians will be familiar with Jeffreys's cogent reasons for and analysis of the dependence of odds ratios on the sample size and again, no paradox. Further, Spiegelhalter requests examples of the use of significance tests in checking scientific theories. The hypothesis of no effect, mentioned in our paper is encountered so frequently that there is no need to publish a list of cases. Also, some theories, for example Milton Friedman's theory of the consumption function predict that parameters will assume particular values and they have been tested extensively in the

literature, many times using inadequate testing methodology. For example, there is much confusion about what significance level to employ when the sample size is large, say about 5,000 as in survey data. With such large samples, empirical workers lament that everything looks significantly different from zero at the 5 percent level. Many of them know that they should not be using the 5 percent level but do not know how to adjust it. Some resort to use of p-values which they find hard to interpret. A posterior odds ratio approach provides a clear-cut solution to these problems given that the prior assumptions employed are deemed satisfactory and other subject matter complications are not present —see Jeffreys (1967, pp. 435-436).

With respect to Spiegelhalter's point regarding accepting H_0 when the likelihood is concentrated around the true parameter value lying $0(n^{-1/2})$ from H_0, we agree with him that for large n "the error is irrelevant" and thus question his charge of "to wrongly accept". Also, as many of our discussants and we noted, our prior distributions under alternative hypotheses are informative, not uninformative as stated by Spiegelhalter. They do, however have the property that if the sample evidence violently conflicts with the null hypothesis, posterior distributions for the parameter or parameters under the alternative hypothesis will be very close to what is obtained with a diffuse prior in estimation, a dove-tailing of Jeffreys's testing and estimation results.

On the dependence of our prior on the sample design, this is not unusual. It is also a feature of the Jeffreys, Box-Tiao, Lindley-Bernardo, Zellner and some other priors. Since information in designing an experiment may not be independent of information about parameters's values, such dependence is reasonable. Also, as Box mentioned at this conference session, uniformative and informative are relative terms, relative to the experiment being considered and thus a dependence between prior and design is not unreasonable. In the case of our multivariate Cauchy prior, it can be interpreted as a standard multivariate Cauchy distribution for standardized regression coefficients much like usual beta coefficients. In the case of one independent variable in a regression, the standardized regression coefficient is precisely the unitless quantity $s_x\beta/\sigma$, where s_x is the sample standard deviation of the independent variable, compatible with and a slight generalization of Jefreys's use of λ/σ in the normal mean problem.

In connection with Spiegelhalter's means problem, since the null hypothesis is equality of means, perhaps reflecting prior information that they may not be far different, it is surprising to see that his prior under the alternative has the means uniformly (over the entire real line?) and independently distributed. This prior implies quite strongly that the means may have widely different values and could help to explain Spiegelhalter's problem. In any event, we did not analyze this problem in our paper. For a sensible analysis of the hypothesis of equality of two means with unequal numbers of observations on each, based on Cauchy priors under the alternative hypothesis and with an application to real data, see Jeffreys (1967, p. 278 ff.).

On the issue of the multiplier for $p_1 - p_0$, as Table 2.1 in our paper referring to the case $p_1 - p_0 = 1$ shows, the multipiier $1n\ (n\text{-}1)$ behaves very reasonably for large n. Also, on choice of models in relation to a loss structure, it is sometimes appropriate to have the loss structure depend on n, as Geisser points out in his comments and this will necessitate a broadened discussion of "the" appropriate multiplier.

In his comments, Geisser describes a frequently encountered circumstance in which investigators are empirically fitting relations with no laws and little or no subject matter theory available. The importance of laws and subject matter theory in science cannot be doubted. But what is one to do in the case described by Geisser? A "starting point" suggested by Jeffreys and others is to consider all variation random until shown otherwise. The hypothesis of "no effect" is thus central as for example in attempting to use a variable to predict stock price changes or gold price changes or in testing a new drug's possible effect. An odds ratio approach seems very appropriate for important problems like these. As regards the Geisser-Eddy predictive scheme, that it provides results that are approximately equivalent to a Bayesian procedure with "penalties (costs or prior weights) that depend on the sample size and the kind of selection error incurred" is very interesting. The afore-mentioned intimate relation of posterior odds ratios and predictive densities, well known to Geisser helps to explain this result. In small samples, however adding too many predictor variables can certainly be harmful in prediction. As the sample size grows, there is a danger that because there is no secure scientific basis for the relationship, it may not be stable. Thus we are back to the desirability of using subject matter theory and laws. On the problem of selecting variables in regression, we have applied the analysis in our paper to the Hald data, also analyzed in the cited Geisser-Eddy paper. We obtained an ordering of models not far different from that of Geisser and Eddy and that based on the residual mean square error criterion. Our results include posterior probabilities for each of the 15 possible models and associated odd ratios. As mentioned at the end of our paper, posterior probabilities have a clear-cut interpretation and can be used to average predictions from alternative models, which may be useful in certain cases and can rationalize *ad hoc* schemes for combining forecasts from alternative models which have appeared in the literature.

With respect to Dickey's remarks, we are at a loss to understand his emphasis on "magic words" and on "automatic procedures which will be universally accepted" in view of our statements regarding prior distributions made on the first page of our paper. Above, we have explained the rationale for the use of our particular Cauchy priors and thus no further comment is needed. Since Jeffreys and we parametrized the normal mean problem in terms of $\eta = \mu/\sigma$ and σ (in Dickey's notation), his equation (8) is relevant and indicates no conlfict between the priors for σ under the null and alternative hypotheses. With respect to other point hypotheses, e.g. $\mu = \mu_0$, at the end of our paper we suggested implicitly that it is possible to write, H_1: $w_i = \epsilon_i$ and H_2: $w_i = \lambda + \epsilon_i$, where $w_i \equiv y_i - \mu_0$ and to proceed to compute the posterior odds ratio for $\lambda = 0$ vs. $\lambda \neq 0$,using Jeffreys's results without difficulty.

Dempster rejects the use of mechanical tail area testing procedures as we do too. He suggests the use of likelihood ratios. For two simple hypotheses, it is well known that the Bayes factor is equal to the likelihood ratio, while for non-simple hypotheses it is equal to a ratio of averaged likelihood functions. Dempster suggests use of the posterior distribution of the likelihood ratio in testing without providing a clear-cut rationale for his procedure. Is the posterior distribution of the likelihood ratio more fundamenta ly linked to relative degrees of confidence in competing hypotheses than is the posterior odds ratio? We believe that it is not even though we find the posterior

distribution of the likelihood ratio interesting.

We agree with Good that his "Device of Imaginary Results" is very important. As we noted, Jeffreys used it, without naming it, in the normal mean problem (and many others) to deduce surprising results associated with the possible use of a normal prior for λ. On "Good's Paradox", it is our opinion that it is reflected in Jeffreys's (1967, p. 255) work.

In closing, we thank the discussants for their comments and hope that our responses help to provide a better understanding of the issues which they have raised.

J.M. BERNARDO (*Universidad de Valencia*):

I am most grateful to all discussants for their thought provoking comments. In the following I shall try to answer their queries.

I certainly agree with Professor Jaynes in considering the determination of reference priors a top priority research problem of Bayesian Statistics, and I am obviously flattered that a physicist with a through understanding of statistics finds my result 'a beautifully neat expression with a clear ring of truth to it'. I object however to his description of my derivation as 'chopping away the prior probability of the null until is reduced to what I consider reasonable'. Indeed this is a mathematical consequence of the procedure; but this is obtained from a well defined general theory on reference distributions which has been shown to work in very different situations. I do not need to invent any *ad hoc* procedures, (like Jeffreys-Zellner-Siow do when they arbitrarily choose a Cauchy prior), but I determine the prior which describes the situation in which most remains to be learned from the experiment, and claim that this is a sensible reference point for scientific inference.

This reference prior is *not* a description of the scientist's beliefs, but a description of the situation in which the experiment could conceivably provide more information on the quantity of interest; no wonder that this might depend on the design of the experiment.

Similarly, I do *not* think the procedure consists of a 'mutilation of equations originally designed to answer Q_1, so as to force them to answer instead Q_2'. Indeed, one must specify what it is considered to be the interesting question, i.e., the quantity of interest in my own terminology. If θ were the quantity of interest I would obtain a reference posterior density $\pi(\theta \mid D)$ for θ. If the question of interest is whether $\theta = \theta_0$ or not. I would obtain a reference posterior probability for $H_0 : \theta = \theta_0$. I dealt with the first question in Bernardo (1979b) and I have tried here to solve the second.

I was very interested in the nearly one-to-one relationship (but for the sign of $\hat{\theta}$) between my reference posterior probability and Laplace's tail area. Indeed, I agree that often the question of interest is whether $\theta > \theta_0$ or not; the corresponding refernce posterior probability is provided in equation (2); see also Bernardo (1979b) in reply to Dawid. However, I do not think that this is the *only* interesting question. I feel it is often convenient in applied work to be able to give a probabilistic description of the plausibility of a sharp null. Confidence levels do *not* have such an interpretation, but reference posterior probabilities do.

Dr. Spiegelhalter wonders what are the appropriate practical situations in which I would use this approach. We all know of those consulting situations in which you are

specifically asked to help some people to perform some or other classical test. As a matter of principle, I refuse to do such a thing, but often do not have the time to go on a lengthy full Bayesian analysis. I would then give these people the reference posterior probability of the hypothesis they wanted to test.

About Definition 6, I do not think it is a little forced; for it is a consequence of the fact that, in the present context, the quantity of interest is *not* θ but, say, $\psi = \psi(\theta)$ defined as $\psi = \psi_0$ if $\theta = \theta_0$ and $\psi = \psi_1$ if $\theta \neq \theta_1$ and, thus, we want to maximize the missing information about ψ, *not* that about θ.

I have not yet had time to extend these results to the general linear model. I would very much like however to see the details of Smith & Spiegelhalter method applied to the particular example I discuss. Informal discussion with Professor Smith suggests that both results are numerically very close.

I certainly agree with Professor Geisser that the question of interest is often prediction. If this is the case, one could obtain the appropriate reference predictive distribution: see Bernardo (1979b) in reply to D.J. Bartholomew; no need, I believe, for Bayes-non Bayes compromises. I do not think however that prediction is the *only* possible question of interest. As in the example given by Professor Jaynes, Science often finds it convenient to work in terms of the statistical falsification of new 'simple' working hypothesis.

Professor Dempster finds it difficult to commit belief to a "prior" distribution derived from an information-theoretic principle; we are not arguing however that one should do so. Indeed, we only consider reference priors as technical tools to produce posteriors which are as little affected as possible, in an information-theoretical sense, by prior opinions. On the other hand, we believe that the mildness of the disparity between those Bayesian techniques and some standard non-Bayesian practice is more than a logical curiosity: indeed, some of those classical techniques have been succesfully used in practice, and we would like to understand why, from a *coherent, unified viewpoint*.

Professor Dempster recognizes the need for Bayesian procedures which provide rational choices between sharp nulls and higher dimensional alternatives and its main use as warning signals for modellers; he provides *no* argument however against the use of reference posterior probabilities with such purpose.

It has been said in this Conference that everything is in Jeffreys. Maybe we have to add 'and/or in Good'. Indeed, I am flattered to discover that the numerical outcome of my well-defined procedure is consistent with the rough and ready rule suggested by Professor Good's remarkable intuition.

I do not think it is sensible to assume $n = 1$ as Professor Akaike does. By so doing he misses the main point of the discussion, namely the behaviour of the proposed procedures as n increases. One may certainly take $n = 1$ if one chooses to call x the vector $x = \{x_1,\ldots,x_n\}$ but then, of course, his argument does not follow. Alternatively one could study the result of using sequentially Akaike's prior: I presume you end up again with Lindley's (or Good's) paradox.

Professor Lindley is certainly right when he mentions the need to think about the real world in order to assess proper prior distributions allowing a subjective Bayesian analysis. I am convinced however that such an analysis is difficult to accept by the

scientific community unless it is accompanied by some *reference* result, conditional only to model and data, with which it could be compared. I have tried to provide such a reference for standard problems of hypothesis testing.

Dr. O'Hagan wonders how would one choose among the different limiting processes one can imagine in (9); I think this is bound to depend on the sort of approximation one is interested in. For, (9) is an *exact* expression, which gives the reference posterior probability of the null when $p(\mu|H_1)=N(\mu|\mu_1,\sigma_1^2)$. The status of equation (17) is however very different from that of (11); while (11) is obtained from an *approximation* to the exact expression (9), valid under certain conditions, (17) is another *exact* expression, which gives the reference posterior probability of the null when no distributional assumptions under the alternative are made.

Professor Zellner mentions once more the dependence of the reference prior on the form of the likelihood function, a feature which is common to most approaches to the problem, including his own. I certainly agree with him on the inevitability of this dependence. Professor Lindley's position was recently made explicit in his contribution to the discussion of Bernardo (1979b).

On Professor Zellner's second point, I certainly do *not* regard as disturbing the fact that $\pi(H_0|D)$ has an upper limit. Indeed, I agree with Professor Jaynes when he questions the need for a probabilistic justification for the maintenance of the *status quo*. The mathematical expression of the fact that, in the absence of evidence against the null, the scientist does not reject H_0, but he is *not* prepared to swear it is true, is the oscillation of $\pi(H_0|D)$ about 1/2, which we obtain under those conditions. I find this far more reasonable than to expect a convergence to one of $\pi(H_0|D)$.

REFERENCES IN THE DISCUSSION

AKAIKE, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30**, 9-14.

ATKINSON, A.C. (1978). Posterior probabilities for choosing a regression model. *Biometrika* **65**, 39-48.

COCCONI, G. and SALPETER, E.E. (1958). *Nuovo Cimento* **10**, 646.

COX, R.T. (1946). *Amer. J. Phys.* **14**, 1-13.

— (1961). *The Algebra of Probable Inference*. Baltimore: John Hopkins.

— (1978). Of inference and inquiry. In *The Maximum-Entropy Formalism*. (Levine, R.D. & Tribus, M. eds.) 119-167. Cambridge, Mass.: M.I.T. Press.

DEMPSTER, A.P. (1973). The direct use of likelihood for significance testing. *Proceedings of the Conference on Foundational Questions in Statistical Inference*. (Barndorff-Nielsen, O., Blaesild, P. and Schou, G., eds.) 335-352. University of Aarhus.

DICKEY, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.* **42**, 204-223.

— (1978). Approximate coherence for regression model inference with a new analysis of Fisher's Broadbalk Wheatfield example. *Bayesian Analysis in Econometric and Statistics: Essays in Honor of Harold Jeffreys*. (Zellner, A. ed.) 333-354. Amsterdam: North-Holland.

GEISSER, S. (1980). The contributions of Sir Harold Jeffreys to Bayesian Inference. In

Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys. (Zellner, A., ed.) 13-20. Amsterdam: North Holland.

GEISSER, S. and EDDY, W.F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74**, 153-160.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence.* New York: Haffners.

— (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28**, 861-880.

— (1969). What is the use of a distribution? In *Multivariate Analysis II.* (Krishaiah, P.R., ed.) 183-203. New York: Academic Press.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/Non Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

JAYNES, E.T. (1976). Confidence intervals vs. Bayesian intervals (with discussion). In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* (Harper, W.L. & Hooker, C.A., eds.) 175-257. Dordrecht, Holland: D. Reidel.

— (1979). Marginalization and prior probabilities (with discussion). In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys.* (Zellner, A., ed.) 43-87. Amsterdam: North-Holland.

JEFFREYS, H. (1963). Review of L.J. Savage, et. al. *The Foundations of Statistical Inference,* (1962). *Technometrics* **5**, 407-410.

— (1967). *Theory of Probability.* (3rd rev. ed.). Oxford: University Press.

— (1979). Personal communication.

KLEIN, F. (1939). *The Monist* **39**, 350-364.

LINDLEY, D.V. (1957). A statistical paradox. *Biometrika* **44**, 187-192.

— (1961). The use of prior probability distributions in statistical inference and decision. *Proc. 4th. Berkeley Symposium* **1**, 453-468.

— (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint: Part 2, Inference.* Cambridge: University Press.

MOLINA, E.C. (1963). Some comments on Bayes' Essay. In *Two Papers by Bayes.* (Deming, W.E., ed.) 7-12. New York: Hafner Pub. Co.

PEARSON, E.S. (1962). Some thoughts on statistical inference. *Ann. Math. Statist.* **33**, 394-403.

SAVAGE, L.J (1962). *The Foundations of Statistical Inference: A Discussion.* London: Methuen.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

SHERWIN, C.W., *et al.* (1960). *Phys. Rev. Lett.* **4**, 399-400.

SMITH, A.F.M. and SPIEGELHALTER, D.J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. B.* **42**, 213-220.

WEISSKOPF, V.F. (1961). Selected topics in theoretical physics. In *Lectures in Theoretical Physics,* **3**, 54-105. (Brittin, W., *et al.,* eds.) New York: Interscience Publishers, Inc.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

— (1977). Maximal data information prior distributions. In *New Developments in the*

Applications of Bayesian Methods. (Aykac, A. and Brumat, C., eds.) Ch. 12, 211-232. Amsterdam: North-Holland.

ZELLNER, A. and SIOW, A. (1979). On posterior odds ratios for sharp null hypotheses and one-sided alternatives. *Tech. Rep.* University of Chicago.