J. R. Statist. Soc. B (1979),
41, *No.* 2, *pp.* 113–147

Reference Posterior Distributions for Bayesian Inference

By Jose M. Bernardo[†]
Universidad de Valencia and Yale University

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 6th, 1978, Professor J. F. C. Kingman in the Chair]

Summary

A procedure is proposed to derive reference posterior distributions which approximately describe the inferential content of the data without incorporating any other information. More explicitly, operational priors, derived from information-theoretical considerations, are used to obtain reference posteriors which may be expected to approximate the posteriors which would have been obtained with the use of proper priors describing vague initial states of knowledge. The results obtained unify and generalize some previous work and seem to overcome criticisms to which this has been subject.

Keywords: non-informative priors; vague initial knowledge; operational priors; information theory; marginalization paradox; stein's paradox; fieller-creasy problem.

1. Introduction

Coherence requirements lead one to believe that, given a sampling model, the only sensible way to make inferences about its parameters is to assess a prior distribution describing one's initial knowledge about their values and to use the data to derive, *via* Bayes' theorem, the appropriate posterior distribution (see, for example, Lindley, 1971, and references therein).

To some statisticians, the obvious dependence of the results on the prior distribution is somewhat disturbing. A possible solution to this difficulty, suggested by Dickey (1973), is to require that a scientific report should display the functional dependence of the posterior distribution on the choice of the prior, for a broad enough range of choices. Among those choices, one would like to include a prior which roughly describes a situation in which little relevant information is available, if only because the resulting *reference* posterior distribution would provide a standard to which other distributions could be referred in order to assess the relative importance of the initial knowledge in the final results.

Much work has been done to formulate prior distributions which add little information to the sample information; this goes back to the early work of Bayes (1763) and Laplace (1825) based on the principle of insufficient reason. Modern approaches to this problem are often based on different types of invariance requirements, as those of Jeffreys (1946, 1939/67), Perks (1947), Barnard (1952), Hartigan (1964, 1965), Stone (1965, 1970), Villegas (1971, 1977a, b), Box and Tiao (1973, Section 1.3), Piccinato (1973, 1977) and Jaynes (1978). Other approaches include the use of limiting forms of conjugate priors as in Haldane (1948), Novick and Hall (1965), Novick (1969) and DeGroot (1970, chapter 10), and different forms of information-theoretical arguments as those of Lindley (1961), Jaynes (1968), Good (1969), Kashyap (1971), Zellner (1971, pp. 51–53, 1977), Bernardo (1975) and Akaike (1978). Moreover, although not directly concerned with the specification of reference priors, results on the conditions for numerical equivalence between classical and Bayesian inference, as those contained in Lindley (1958, 1965), Welch and Peers (1963), Geisser and Cornfield (1963) and Bartholomew (1965), are often relevant for its discussion.

[†] Now at Departamento de Bioestadística, Facultad de Medicina, Ave. Blasco Ibáñez 17, Valencia-10, Spain.

However, although we have many results which provide seemingly appropriate reference priors for a number of inference problems, no general theory has emerged which is capable of dealing with them all. More important, however, is that none of the procedures so far proposed seem to be able to deal with a number of serious criticisms raised against the uncritical use of (usually improper) reference priors. These criticisms include the inadmissibility results of Stein (1956), the marginalization paradoxes of Dawid *et al.* (1973), the results on strong inconsistency of Stone (1976) and Stein's paradox on the sum of squares of normal means (Stein 1959; Efron, 1973; Cox and Hinkley 1974, p. 383), and clearly apply as well to fiducial and to structural inference.

This paper is an attempt to overcome these difficulties and suggest an operative procedure to derive reference posterior distributions which approximately describe the kind of inferences which one is entitled to make with little relevant initial information. The approximation referred to is to be taken in the sense of Dickey (1976); indeed, a real situation in which little initial information is available will be modelled by an *operational* (often improper) reference prior in such a way that the resulting reference posterior may be expected to approximate the posterior which would have been obtained with the use of a proper prior describing such vague initial knowledge. With expressions like "little initial information" or "vague initial knowledge" we intend to describe a situation in which most remains to be learned from the data, in a sense to be made precise.

We shall conclude that the relevant reference prior may differ according to the parameter of interest. Thus, the operational prior used to derive the reference posterior for a normal mean turns out to be different from that required to obtain a reference posterior for the coefficient of variation. This was only to be expected, since vague initial information about the mean approximates a different state of knowledge from vague initial information about the coefficient of variation, and should therefore be modelled by a different function.

People have sometimes questioned the need for reference distributions. We find it difficult however to avoid the need for an origin from which to measure precisely the relevance of the initial information. Particularly in scientific work, it seems difficult to deny the convenience of the eventual availability of standard posterior distributions which do not incorporate the scientist's personal opinions. The point was argued in Novick (1969) and ensuing discussion.

Nevertheless, although the proposed operational priors depend on the likelihood function, we claim that their use as technical tools to obtain reference posteriors which provide origins for admissible inferences is compatible with a subjective view of probability. Here, and in the rest of the paper, we mean by admissible inferences those which may be produced, via Bayes' theorem, with a proper prior compatible with whatever "objective" knowledge one is willing to assume. Indeed, a reference posterior may be seen as an approximation to the personal posterior which would have been obtained by someone who happened to have little initial information; a Bayesian statistician with a subjective prior could presumably be interested in comparing his own posterior with the reference posterior obtained by his uninformed colleague.

In this paper we intend only to provide a heuristic discussion of the basic ideas underlying our construction of reference posterior distributions to see whether they are sound; we feel that, at this point, much attention to mathematical detail would be premature.

In the next section, some notation is introduced and the procedure to derive reference posterior distributions is described. In Section 3, their behaviour is investigated in a number of examples; in particular, it is proved that in the finite discrete case our result coincides with Jaynes' solution (1968) and that in the one-dimensional continuous case, under regularity conditions, Jeffreys' prior is obtained.

Section 4 deals with the general situation in which nuisance parameters are present and offers some examples. In Section 5 it is found that, with this formulation, marginalization paradoxes do not seem to appear; moreover, reference posterior distributions are obtained for inference problems which have been regarded as somewhat controversial. These include

Stein's paradox on the sum of the squares of normal means and the Fieller–Creasy problem on the ratio of normal means.

Finally, we consider in the last section the limitations of the proposed procedure and suggest areas for additional research.

2. REFERENCE DISTRIBUTIONS

Let us assume that the objective of a piece of research is to improve one's knowledge about some parameter of interest θ belonging to a parameter space Θ. Let $\varepsilon = \{X, \Theta, p(x\mid\theta)\}$ be the experiment which consists of one observation of the random quantity $x \in X$ which is distributed, for some $\theta \in \Theta$, according to the probability density $p(x\mid\theta)$ with respect to some σ-finite dominating measure on X. Without loss of generality, we shall assume that the probability densities of x which correspond to different values of θ differ at least on a set of non-zero (dominating) measure. Reference will often be made to the experiment $\varepsilon(k)$ which consists of k independent replications of ε, each with the same value of θ.

For simplicity in notation, we shall not generally attempt to be specific in describing the density functions. Thus, $p(x)$ will denote the density function of the random quantity x and $p(\theta)$ that of the random quantity θ without any suggestion that the random quantities x *and* θ have the same distribution. Specific densities, used to construct examples, will be denoted by specific symbols. Thus, if θ has a Beta distribution with parameters a and b, its density function will be denoted by $\mathrm{Be}(\theta\mid a,b)$, where

$$\mathrm{Be}(\theta\mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1}.$$

Let $p(\theta)$ be a prior probability density of θ with respect to some dominating measure on Θ. Without loss of generality, assume $p(\theta)$ strictly positive, i.e. such that $p(\theta)>0$ for all $\theta\in\Theta$. Following Lindley (1956), the *expected information* about θ to be provided by $\varepsilon = \{X, \Theta, p(x\mid\theta)\}$ when the prior density of θ is $p(\theta)$ is defined to be

$$I^{\theta}\{\varepsilon, p(\theta)\} = \int p(x) \int p(\theta\mid x) \log\frac{p(\theta\mid x)}{p(\theta)}\,d\theta\,dx, \tag{1}$$

where $p(x) = \int p(x\mid\theta)p(\theta)\,d\theta$ and $p(\theta\mid x) = p(x\mid\theta)p(\theta)/p(x)$.

It is worth pointing out that the amount of information defined by (1) does not depend on the dominating measures and may be expressed directly in terms of Radon–Nikodym derivatives as

$$I^{\theta}\{\varepsilon, P(\theta)\} = \int\int dP(\theta\mid x)\,dP(x) \log\frac{dP(\theta\mid x)}{dP(\theta)}.$$

However, for the sake of simplicity, we shall be using the definition in density form with either the Lebesgue or the counting measures as dominating measures.

Although other measures of information have been proposed in the literature, the logarithmic measure defined above seems clearly preferable to us, both in terms of its properties: invariance, non-negativity, concavity; see Lindley (1956), and in terms of its axiomatic justification: Shannon (1948) and Lee (1964) for the discrete case; Good (1966) for a probabilistic explanation of information; Bernardo (1979) for a general decision-theoretical argument.

The basic idea underlying the construction of a reference posterior may now be stated as follows. Consider the quantity $I^{\theta}\{\varepsilon(k), p(\theta)\}$, i.e. the amount of information about θ to be expected from k independent replications of ε, and let C be the class of admissible priors, i.e. those compatible with whatever agreed "objective" initial information one is willing to assume. By performing infinite replications of ε one would get to know precisely the value of θ. Thus, $I^{\theta}\{\varepsilon(\infty), p(\theta)\}$ measures the amount of missing information about θ when the prior is $p(\theta)$. It seems natural to define "vague initial knowledge" about θ as that described by the density $\pi(\theta)$ which maximizes the missing information in the class C. The reference posterior distribution for θ after x has been observed, to be denoted $\pi(\theta|x)$, may now be obtained *via* Bayes' theorem so that $\pi(\theta|x) \propto p(x|\theta)\pi(\theta)$.

In the continuous case, it is usually true that $I^\theta\{\varepsilon(\infty), p(\theta)\} = +\infty$, for all $p(\theta)$. This is to be expected since an infinite amount of information would be required to know exactly a real number. However, one may define a reference posterior as a limiting result. By $\lim p_k(\theta) = p(\theta)$ we mean that the corresponding sequence of distribution functions converges to the distribution function of the limit in all its points of continuity. We shall assume that the class C of admissible priors is compact with respect to the topology induced by such convergence.

Definition 1. Let x be the result of an experiment $\varepsilon = \{X, \Theta, p(x|\theta)\}$ and let C be the class of admissible priors. The reference posterior of θ after x has been observed is defined to be $\pi(\theta|x) = \lim \pi_k(\theta|x)$, where $\pi_k(\theta|x) \propto p(x|\theta)\pi_k(\theta)$ is the posterior density corresponding to that prior $\pi_k(\theta)$ which maximizes $I^\theta\{\varepsilon(k), p(\theta)\}$ in C. A reference prior for θ is a positive function $\pi(\theta)$ which satisfies $\pi(\theta|x) \propto p(x|\theta)\pi(\theta)$.

The compactness requirement for C is necessary to guarantee the existence of the maxima involved in the definition. Since I^θ is concave as a functional of $p(\theta)$ (Lindley, 1956) these maxima will be unique. If the class of admissible priors is not compact one could construct an expanding sequence of compact sets converging to C, derive the corresponding sequence of reference posteriors using Definition 1, and define its limit to be the appropriate reference density.

It may seem unnecessarily complicated to define $\pi(\theta)$ indirectly using the limiting process in the sequence of posteriors. However, a direct definition in terms of $\pi(\theta) = \lim \pi_k(\theta)$ entails difficulties. For instance, with a sequence of priors $\pi_k(\theta) = \mathrm{Be}(\theta|1/k, 1/k)$ the limit of the corresponding sequence of posteriors after observing r successes in n Bernouilli trials with parameter θ would be $\pi(\theta|r) = \mathrm{Be}(\theta|r, n-r)$, implying an operational prior $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$; however, with the topology adopted, $\lim \pi_k(\theta)$ is the discrete distribution $\pi(\theta = 0) = \pi(\theta = 1) = \frac{1}{2}$.

Very often, under regularity conditions, a reference prior may be obtained much more rapidly than Definition 1 may suggest. For, if $\mathbf{z} = \{x_1, \ldots, x_k\}$ is the result of $\varepsilon(k)$, we may write

$$I^\theta\{\varepsilon(k), p(\theta)\} = \int p(\mathbf{z}) \int p(\theta|\mathbf{z}) \log \frac{p(\theta|\mathbf{z})}{p(\theta)}\, d\theta\, d\mathbf{z}$$

$$= H\{p(\theta)\} - \int p(\mathbf{z})\, H\{p(\theta|\mathbf{z})\}\, d\mathbf{z}, \tag{2}$$

where $H\{p(\theta)\} = -\int p(\theta) \log p(\theta)\, d\theta$ is called the *entropy* of $p(\theta)$ for historical reasons. Using $p(\mathbf{z}) = \int p(\mathbf{z}|\theta) p(\theta)\, d\theta$ and reversing the order of integration in (2) we have

$$I^\theta\{\varepsilon(k), p(\theta)\} = H\{p(\theta)\} - \int p(\theta) \int p(\mathbf{z}|\theta)\, H\{p(\theta|\mathbf{z})\}\, d\mathbf{z}\, d\theta$$

$$= \int p(\theta) \log \left\{ \exp\left(- \int p(\mathbf{z}|\theta)\, H\{p(\theta\ \mathbf{z})\}\, d\mathbf{z} \right) \Big/ p(\theta) \right\} d\theta \tag{3}$$

and, also,

$$I^\theta\{\varepsilon(k), p(\theta)\} = H\{p(\theta)\} + \int p(\theta) \int p(\mathbf{z}|\theta) \log p(\theta|\mathbf{z})\, d\mathbf{z}\, d\theta$$

$$= \int p(\theta) \log \left\{ \exp\left(\int p(\mathbf{z}|\theta) \log p(\theta|\mathbf{z})\, d\mathbf{z} \right) \Big/ p(\theta) \right\} d\theta. \tag{4}$$

The equivalent expressions (3) and (4) are both of the form $\int p(\theta)\log\{f(\theta)/p(\theta)\}\,d\theta$, which is maximized (provided $\int f(\theta)\,d\theta<\infty$) when $f(\theta)\propto p(\theta)$ as an elementary exercise in calculus of variations shows. Thus, under regularity conditions to guarantee the operations involved, two sequences of prior distributions approaching the reference prior, in the sense of Definition 1, are approximately provided by

$$\pi_k(\theta)\propto \exp\left(-\int p(\mathbf{z}\mid\theta)\,H\{p^*(\theta\mid\mathbf{z})\}\,d\mathbf{z}\right) \tag{5}$$

and

$$\pi_k(\theta)\propto \exp\left(\int p(\mathbf{z}\mid\theta)\log p^*(\theta\mid\mathbf{z})\,d\mathbf{z}\right) \tag{6}$$

for large values of k, where $p^*(\theta\mid\mathbf{z})$ is the asymptotic posterior density of θ, which is independent of the prior.

It may be noted that, as one would require, the results of (5) or (6) are not affected if the data $\mathbf{z}$ are replaced by a sufficient statistic $t=t(\mathbf{z})$. Indeed, their common limiting result in the sense of Definition 1, the reference prior $\pi(\theta)$, will not even be affected if the data $\mathbf{z}$ in (5) or (6) are replaced by an asymptotically sufficient statistic, that is by some function $t=t(\mathbf{z})$ such that, as $k\to\infty$, $p(\theta\mid\mathbf{z})=p(\theta\mid t)\{1+o(1)\}$ uniformly.

Moreover, as a consequence of the invariance of I^θ under one-to-one transformations of θ the procedure is invariant under reparametrization. This is trivial in the discrete case for then reparametrization reduces to a relabelling which does not affect the probabilities. If θ is continuous and $\zeta=\zeta(\theta)$ is a one-to-one transformation of θ, a sequence of priors approaching the reference prior for ζ is

$$\begin{aligned}\pi_k(\zeta)&\propto \exp\left(\int p(\mathbf{z}\mid\zeta)\log p^*(\zeta\mid\mathbf{z})\,d\mathbf{z}\right)\\ &=\exp\left(\int p(\mathbf{z}\mid\theta)\log\{|J|\,p^*(\theta\mid\mathbf{z})\}\,d\mathbf{z}\right)\\ &=|J|\exp\left(\int p(\mathbf{z}\mid\theta)\log p^*(\theta\mid\mathbf{z})\,d\mathbf{z}\right)=|J|\,\pi_k(\theta),\end{aligned} \tag{7}$$

where $|J|=|\partial\theta/\partial\zeta|$ is the Jacobian of the transformation. Thus, as one would require, the reference prior for a one-to-one transformation of θ may be obtained from that of θ by the appropriate change of variable.

3. SOME EXAMPLES

3.1. *The Finite Discrete Case*

If θ may only take a finite number of values (say m) then, for any experiment ε, the reference prior in the unrestricted class of all probability distributions of θ is the uniform distribution $\pi(\theta)=\{1/m,\ldots,1/m\}$. For, Rényi (1964) showed that, in the discrete finite case, $\lim_{k\to\infty}H\{p(\theta\mid\mathbf{z})\}=0$ and thus, using (5), we have $\pi(\theta)\propto 1$.

More generally, using (2), we obtain that in the finite discrete case, the missing amount of information is precisely the original Shannon entropy, i.e. $I^\theta\{\varepsilon(\infty),p(\theta)\}=H\{p(\theta)\}$. One may note that the Shannon entropy was axiomatically developed as a measure of uncertainty in the finite discrete case. We see that the concept of missing information contains this as a particular case. As a consequence, the reference prior in a given class C is here that which maximizes the entropy in such a class. This agrees with Jaynes (1968).

The infinite discrete case cannot be handled in a similarly easy way because no general results seem to be available on the asymptotic posterior entropy of a discrete variable which may take an infinite number of values. However, the problem may usually be solved by embedding the model in a continuous one for which such type of results do exist (see Section 3.3).

3.2. *The General Continuous Case*

Under regularity conditions, the limiting form of (5) and (6) takes a very simple form. For, if a maximum likelihood estimate $\hat{\theta} = \hat{\theta}(\mathbf{z})$ exists, the asymptotic posterior distribution $p^*(\theta|\mathbf{z})$ usually depends only on the data through $\hat{\theta}$. Thus, the asymptotic posterior entropy may be written as

$$\begin{aligned} H\{p^*(\theta|\mathbf{z})\} &= -\int p^*(\theta|\,\hat{\theta})\log p^*(\theta|\,\hat{\theta})\,d\theta \\ &= -\log p^*(\hat{\theta}|\,\hat{\theta}) + o(1) \\ &= K(\hat{\theta}) + o(1) \end{aligned} \tag{8}$$

where $K(\hat{\theta}) = -\log p^*(\hat{\theta}|\,\hat{\theta})$, since for large k the posterior density will concentrate around $\hat{\theta}$. Moreover, since for large k the likelihood $p(\mathbf{z}|\,\theta)$ will also concentrate around its maximum $\hat{\theta}$, we have

$$\int p(\mathbf{z}|\,\theta)\,K(\hat{\theta})\,dz = K(\theta) + o(1)$$

so that both equations (5) and (6) become

$$\pi_k(\theta) \propto \exp\{-K(\theta)\}\{1 + o(1)\} \tag{9}$$

and the reference posterior density of θ after x has been observed is simply

$$\pi(\theta|\,x) \propto p(x|\,\theta)\exp\{-K(\theta)\}. \tag{10}$$

3.3. *The "Regular" Continuous Case*

Assume the usual regularity conditions for asymptotic normality of the posterior distribution of θ (cf. Lindley, 1961; Walker, 1969; Johnston, 1970; Dawid, 1970) so that $p^*(\theta|\mathbf{z})$ is normal with mean $\hat{\theta}$, the maximum likelihood estimate, and precision (inverse of the variance) $ki(\hat{\theta})$, where

$$i(\theta) = -\int p(x|\,\theta)\frac{\partial^2}{\partial\theta^2}\log p(x|\,\theta)\,dx. \tag{11}$$

It is easily verified that if θ has a normal distribution with mean μ and precision h, its entropy is

$$H\{N(\theta|\,\mu,h)\} = \tfrac{1}{2}\log(2\pi e/h). \tag{12}$$

Using (12), the asymptotic posterior entropy of θ is

$$H\{p^*(\theta|\,\mathbf{z})\} = \tfrac{1}{2}\log(2\pi e/k) - \tfrac{1}{2}\log i(\hat{\theta}) + o(1)$$

so that using (8) and (9) and leaving out an irrelevant constant

$$\pi(\theta) \propto \exp\{-K(\theta)\} \propto i(\theta)^{\frac{1}{2}} \tag{13}$$

which is, of course, Jeffreys' (1946, 1939/67) prior.

Alternative justifications for this prior have been given by Perks (1947), Lindley (1961), Welch and Peers (1963), Hartigan (1965), Good (1969), Kashyap (1971), Box and Tiao (1973, 1.3) and Akaike (1978). From our own approach, Jeffreys is the appropriate reference prior if, and only if, there are no nuisance parameters, and the usual form of asymptotic normality may be guaranteed.

The argument may easily be extended to the multivariate case, so that we obtain Jeffreys' multivariate prior for simultaneous inference about all the parameters. We do not know of any objection to the use of such a prior for simultaneous inferences, i.e. to derive a joint reference posterior. If, however, we are interested in, say, one of the parameters, the rest being nuisance parameters, the situation is quite different, and the appropriate reference prior is no longer Jeffreys' multivariate prior. Indeed the reference prior to obtain a reference posterior for μ in a Normal situation with both parameters unknown is $\pi(\mu, \sigma) \propto \sigma^{-1}$ and *not* Jeffreys' $\pi(\mu, \sigma) \propto \sigma^{-2}$ (see Section 4).

The preceding argument may easily be modified to obtain the reference prior for a quantity θ whose asymptotic posterior distribution is known. If, in particular, the asymptotic posterior distribution of θ is known to be normal with variance $\sigma^2(\tilde{\theta})/k$ which depends on some asymptotically consistent estimate $\tilde{\theta}$ of θ then, by the argument just presented, the reference prior for θ will be $\pi(\theta) = 1/\sigma(\theta)$. This makes precise the conditions under which Perks' (1947) suggestion, based purely on intuitive grounds, is to be used. An interesting application of this result occurs in Stein's paradox about the sum of squares of normal means (see Section 5.3).

3.4. *Binomial Data*

The problem of making inferences about the parameter θ of a binomial distribution has often been regarded as controversial. Suggested reference priors are uniform (Bayes, 1763; Laplace, 1825); $\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$ (Jeffreys, 1946; Perks, 1947) and $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ (Haldane, 1948; Jaynes, 1968; Novick, 1969). Their relative merits are discussed in Jeffreys' book (1939/67, p. 184) and in the discussion following Novick's (1969) paper. It follows from the results in Section 3.3 that our approach leads to Jeffreys'. Thus, if n independent observations are taken from a Bernouilli process with parameter θ, r of which result in successes, our reference posterior would be $\text{Be}(\theta \mid r+\frac{1}{2}, n-r+\frac{1}{2})$. In particular, if $r = 0$, we obtain the reference posterior $\text{Be}(\theta \mid \frac{1}{2}, n+\frac{1}{2})$ while the posterior density using Haldane's prior would still be improper. Now consider that a random sample of 60 individuals is checked for lung cancer and none of them has the disease. We would conclude for instance that, in the absence of other sources of information, we are prepared to bet approximately evenly on the proportion of people in the population with lung cancer being less than 0·4 per cent. With Haldane's prior, inferences about θ cannot be made since the posterior is improper; we find this less than adequate.

3.5. *Non-regular Continuous Case*

We shall conclude this section by considering an example in which the asymptotic posterior distribution is not normal. Let $\mathbf{z} = \{x_1, \ldots, x_k\}$ be a random sample from a uniform distribution over the interval $(\theta-\frac{1}{2}, \theta+\frac{1}{2})$ and suppose that we are interested in the value of θ. It may be verified that the asymptotic posterior distribution of θ is uniform over the interval $(x_{\max}-\frac{1}{2}, x_{\min}+\frac{1}{2})$ where $x_{\max}$ and $x_{\min}$ are respectively the maximum and minimum values in the sample. Thus,

$$-H\{p^*(\theta \mid \mathbf{z})\} = \int p^*(\theta \mid \mathbf{z}) \log p^*(\theta \mid \mathbf{z})\, d\theta$$

$$= -\log\{1-(x_{\max}-x_{\min})\} + o(1)$$

and, moreover,

$$-\int p(\mathbf{z} \mid \theta) H\{p^*(\theta \mid \mathbf{z})\}\, d\mathbf{z} = \log \tfrac{1}{2}(k+1) + o(1)$$

which is independent of θ. Thus, using (5), the reference prior for θ is uniform and therefore, using Bayes' theorem, the reference posterior distribution $\pi(\theta \mid \mathbf{z})$ is a uniform distribution over $(x_{\max}-\frac{1}{2}, x_{\min}+\frac{1}{2})$.

4. Nuisance Parameters

Let us consider now the general case in which we want to use the result x of an experiment $\varepsilon = \{X, \Psi, p(x \mid \psi)\}$ to make inferences about some function of the parameter $\theta = \theta(\psi)$ rather than about the parameter ψ itself. Without loss of generality, assume that the quantity of interest $\theta = \theta(\psi)$ consists of the first component of ψ so that $\psi = (\theta, \omega)$, where ω is some nuisance parameter since, otherwise, an appropriate transformation could be made to achieve such a situation.

Extending Lindley's (1956) definition, the expected information about θ to be provided by $\varepsilon = \{X, \Psi, p(x \mid \theta, \omega)\}$ when the prior density of $\psi = (\theta, \omega)$ is $p(\psi) = p(\theta) p(\omega \mid \theta)$ is defined to be

$$I^{\theta}\{\varepsilon, p(\theta, \omega)\} = \int p(x) \int p(\theta \mid x) \log \frac{p(\theta \mid x)}{p(\theta)} \, d\theta \, dx, \tag{14}$$

with $p(x) = \int p(x \mid \theta, \omega) p(\theta, \omega) \, d\theta \, d\omega$, $p(x \mid \theta) = \int p(x \mid \theta, \omega) p(\omega \mid \theta) \, d\omega$ and $p(\theta \mid x) = p(x \mid \theta) p(\theta) / p(x)$. Note that the expected information about θ depends on the entire prior $p(\psi) = p(\theta, \omega)$ and not only on the corresponding marginal $p(\theta)$. It may be shown (Bernardo, 1978) that I^{θ} retains the appealing properties (additivity, non-negativity, etc.) which I^{ψ} has and, furthermore, that for all $p(\theta, \omega)$ one has

$$I^{\psi}\{\varepsilon, p(\theta, \omega)\} = I^{\theta}\{\varepsilon, p(\theta, \omega)\} + \int p(\theta) \, I^{\omega \mid \theta}\{\varepsilon, p(\theta, \omega)\} \, d\theta, \tag{15}$$

where

$$I^{\omega \mid \theta}\{\varepsilon, p(\theta, \omega)\} = \int p(x \mid \theta) \int p(\omega \mid \theta, x) \log \frac{p(\omega \mid \theta, x)}{p(\omega \mid \theta)} \, d\omega \, dx \tag{16}$$

so that, in particular, $I^{\theta} \leqslant I^{\psi}$.

For any given conditional prior $p(\omega \mid \theta)$ on the nuisance parameter, the expected information about θ may be computed from (14) and thus, using the argument in Section 2, a reference prior $\pi(\theta)$ for the parameter of interest may be derived as the limit of

$$\pi_k(\theta) \propto \exp\left(- \int p(\mathbf{z} \mid \theta) \, H\{p^*(\theta \mid \mathbf{z})\} \, d\mathbf{z} \right) \tag{17}$$

where $p(\mathbf{z} \mid \theta) = \int p(\mathbf{z} \mid \theta, \omega) p(\omega \mid \theta) \, d\omega$. A reference posterior distribution for θ may now be obtained by the formal use of Bayes' theorem so that

$$\pi(\theta \mid x) \propto \int p(x \mid \theta, \omega) p(\omega \mid \theta) \pi(\theta) \, d\omega. \tag{18}$$

The reference posterior thus obtained will generally depend on $p(\omega \mid \theta)$.

The conditional prior of the nuisance parameters $p(\omega \mid \theta)$ may be chosen so as to describe personal opinions, previous empirical "objective" knowledge or, alternatively, to describe some form of diffuse opinions about ω given θ, using the procedure described in Section 2. Each of these assessments of $p(\omega \mid \theta)$ will give rise to a different reference posterior distribution for the parameter of interest θ. This battery of reference posteriors would establish different "origins" to make inferences about the parameter of interest depending on the assumptions that one is willing to make about the nuisance parameters.

Occasionally, one may find a conditionally sufficient statistic $t = t(x)$ whose sampling distribution only depends on θ, i.e. such that $p(t \mid \theta, \omega) = p(t \mid \theta)$. By conditionally sufficient we mean that, given $p(\omega \mid \theta)$, the posterior distribution of θ only depends on t, i.e. $p(\theta \mid x) = p(\theta \mid t)$ whatever the prior $p(\theta)$ might be. This is the situation in which the marginalization paradoxes (Dawid *et al.*, 1973) may occur. If t is conditionally sufficient for θ then the reference posterior density of θ, $\pi(\theta \mid x) = \pi(\theta \mid t) \propto p(t \mid \theta) \pi(\theta)$ does *not* depend on the exact form of $p(\omega \mid \theta)$ and may be interpreted as (i) an origin for those inferences about θ from priors for which t is conditionally sufficient or (ii) an origin for those inferences about θ based solely on t, rather than the complete data x, whatever the prior $p(\theta, \omega)$ might be.

It is important to distinguish between the quantity of interest θ and the complete parameter $\psi = (\theta, \omega)$; this, to the best of our knowledge, has not been done previously. We proceed to illustrate the difference by means of some examples.

4.1. *The Counterfeit Coin*

Let us suppose that ε consists of one toss of a coin which is known to be fair ($\psi = \psi_0$) or double headed ($\psi = \psi_1$) or double tailed ($\psi = \psi_2$), and let x be the result of the toss, where $x = 1$ stands for "head" and $x = 0$ for "tail". Thus,

$$p(x|\psi_0) = \tfrac{1}{2}, \quad p(x|\psi_1) = x, \quad p(x|\psi_2) = 1-x, \quad x = 0, 1.$$

Moreover, assume that we are interested on whether the coin is fair or not. We may describe the parameter ψ as $\psi = (\theta, \omega)$, where θ specifies whether the coin is fair ($\theta = \theta_0$) or not ($\theta = \theta_1$) and ω specifies whether the coin is double headed ($\omega = \omega_1$) or double tailed ($\omega = \omega_2$) given that it is not fair. We are interested in a reference posterior distribution for θ.

According to the result stated in Section 3.1, the reference prior for θ is the uniform distribution $\pi(\theta_0) = \pi(\theta_1) = \frac{1}{2}$ whatever the prior for ω might be. Similarly, if we do not have (or do not wish to use) any relevant information about ω given θ, we may use the same argument to obtain the reference prior for ω given θ which, again, will be uniform; indeed, we would need that prior $\pi(\omega|\theta)$ which maximizes the missing information about ω given θ, i.e. $I^{\omega|\theta}\{\varepsilon(\infty), p(\theta, \omega)\} = H\{p(\omega|\theta)\}$. This is maximized by $\pi(\omega_1|\theta = \theta_1) = \pi(\omega_2|\theta = \theta_1) = \frac{1}{2}$. Thus, the operational prior to make inferences about θ is

$$\pi_\theta(\psi) = \pi(\theta)\pi(\omega|\theta) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}). \tag{19}$$

Using Bayes' theorem it is easily established that the corresponding reference posterior for θ after n tosses of the coin, r of which resulted in heads, is

$$\begin{aligned} \pi(\theta_0|r) &= 1/(1+2^{n-1}), \quad \text{if } r = 0 \text{ or } r = n \\ &= 1, \quad \text{otherwise} \end{aligned} \tag{20}$$

and $\pi(\theta_1|r) = 1-\pi(\theta_0|r)$. Inspection shows that (20) behaves as one would expect from a posterior which reflects the inferential content of the data without incorporating any other information. For example, if $n = 1$, then $\pi(\theta_0|r) = \frac{1}{2}$ ($r = 0, 1$) corresponding to the obvious fact that the first toss of the coin gives no information on its own about whether the coin is fair or not and thus, *in the absence of any other source of information*, both possibilities should have the same probability. Note that the uniform prior for ψ, $\pi(\psi) = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ which has often been described as a "universal" representation of ignorance in the discrete case yields, for $n = 1$, $p(\theta_0|r) = \frac{1}{3}$ ($r = 0, 1$) pointing out the fact that although "non-informative" with respect to ψ, the uniform prior $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ describes some information about θ, making twice as likely that the coin is not fair than that it is fair.

A superficial analysis of this example could lead one to think that this approach can justify all sorts of priors on discrete parameters, simply by considering suitable many-to-one transformations. Of course, what we argue is that one should have a reference uniform prior in the discrete case on the parameter of interest, i.e. on that which is the immediate object of inference, regardless of how it may relate to other parameters in the model.

4.2. *Reference Posteriors for the Normal Case*

Let $\mathbf{z} = \{x_1, \ldots, x_k\}$ be a random sample from a normal distribution with mean μ and standard deviation σ, and suppose that we are interested in the value of μ, σ being a nuisance parameter.

It is well known (see, for example, DeGroot, 1970, Section 10.10) that the asymptotic posterior distribution of μ is normal with mean $\bar{x} = \Sigma x_i/k$ and variance $\hat{\sigma}^2/k$ where $\hat{\sigma}^2 = s^2 = \Sigma(x_i - \bar{x})^2/k$. Thus, using (17) and (12), the reference prior for μ will be the limit, in the sense of Definition 1, of

$$\pi_k(\mu) = \exp\left(-\int p(\mathbf{z}\,|\,\mu)\,\tfrac{1}{2}\log\{(2\pi e/k)\,\hat{\sigma}^2\}\,d\mathbf{z}\right), \tag{21}$$

where $p(\mathbf{z}\,|\,\mu) = \int \Pi N(x_i\,|\,\mu,\sigma)\,p(\sigma\,|\,\mu)\,d\sigma$. Thus, since the likelihood will concentrate for large k on its maximum,

$$\pi_k(\mu) \propto \exp\left(-\int p(\sigma\,|\,\mu)\log\sigma\,d\sigma\right) + o(1) \tag{22}$$

so that, as we have anticipated, the reference (marginal) prior of μ will generally depend on $p(\sigma\,|\,\mu)$. If σ is *a priori* considered independent of μ so that $p(\sigma\,|\,\mu) = p(\sigma)$, and only then, the integral (22) will not depend on μ and the reference prior for μ will be uniform.

We may want $p(\sigma\,|\,\mu)$ to describe diffuse opinions about σ given μ. Then, using the argument in Section 2, one would like to maximize the missing information about σ given μ. The same argument used to derive (5) leads then to

$$\pi_k(\sigma\,|\,\mu) \propto \exp\left(-\int p(\mathbf{z}\,|\,\mu,\sigma)\,H\{p^*(\sigma\,|\,\mu,\mathbf{z})\}\,d\mathbf{z}\right). \tag{23}$$

Now, the asymptotic posterior distribution of σ given μ is normal with variance $\hat{\sigma}^2/2k$ so that using (23) and (12),

$$\pi_k(\sigma\,|\,\mu) \propto \exp\left(-\int \Pi N(x_i\,|\,\mu,\sigma)\,\tfrac{1}{2}\log\{(\pi e/k)\,\hat{\sigma}^2\}\,d\mathbf{z}\right)$$

and therefore

$$\pi_k(\sigma\,|\,\mu) \propto \sigma^{-1}\{1 + o(1)\}.$$

Consequently, the joint reference prior to make inferences about μ is

$$\pi_\mu(\mu,\sigma) = \pi(\mu)\,\pi(\sigma\,|\,\mu) = \sigma^{-1},$$

that is the left Haar invariant measure already defended by Jeffreys (1939/67, p. 138), Barnard (1952) and Stone (1965) on different grounds. The corresponding reference posterior for μ is the familiar Student t with $n-1$ degrees of freedom, i.e.

$$\pi(\mu\,|\,x_1,\ldots,x_n) \propto [1 + \{(x-\mu)/s\}^2]^{-\frac{1}{2}n},$$

where $s^2 = \Sigma(x_i - \bar{x})^2/n$.

Similarly, if we are interested in σ, μ being now the nuisance parameter, one may use an analogous argument to obtain $\pi_\sigma(\mu,\sigma) = \sigma^{-1}$ as the reference prior to make inferences about σ. The corresponding reference posterior distribution of σ is

$$\pi(\sigma\,|\,x_1,\ldots,x_n) \propto \sigma^{-n}\exp\{-ns^2/2\sigma^2\},$$

i.e. ns^2/σ^2 has the familiar χ^2_{n-1} distribution.

However, as we shall see in the next section, the reference prior to make inferences about $\lambda = \mu/\sigma$ is no longer σ^{-1} but one that avoids the marginalization paradox discussed by Stone and Dawid (1972).

5. A Solution to some Controversial Problems

5.1. *Marginalization Paradoxes*

Let us suppose that in the normal case discussed in Section 4.2, one is interested in the value of $\lambda = \mu/\sigma$. Then, if one insists on using $\pi(\mu,\sigma) \propto \sigma^{-1}$ as an operational prior, problems arise. For (Stone and Dawid, 1972) the posterior distribution of λ obtained with such a prior depends on the data through the statistic $r = (\Sigma x_i)/(\sqrt{\Sigma x_i^2})$ whose sampling distribution

$$p(r\,|\,\mu,\sigma) = \exp(-\tfrac{1}{2}n\lambda^2)\{1-(r^2/n)\}^{\frac{1}{2}(n-3)} \int_0^\infty \omega^{n-1}\exp\{-\tfrac{1}{2}\omega^2+r\lambda\omega\}\,d\omega$$

only depends on λ. Therefore, one would expect to be able to "match" the original inferences about λ by the use of $p(r\,|\,\lambda)$ together with some appropriate prior for λ. However, no such a prior exists. This type of *marginalization paradox* further explored by Dawid *et al.* (1973) and recently discussed by Jaynes (1978), appears in a large number of multi-parameter problems. This makes it difficult to believe that such a thing as an all-purpose representation of "vague knowledge" about the parameters of a given model is possible.

In a previous paper (Bernardo, 1977b) we applied the procedure described above to derive the reference prior to make inferences about $\lambda = \mu/\sigma$. It turns out that, in terms of λ and σ and whatever the conditional prior $p(\sigma\,|\,\lambda)$ might be, the reference prior for λ is

$$\pi(\lambda) \propto (1+\tfrac{1}{2}\lambda^2)^{-\frac{1}{2}}$$

and that of σ given λ, $\pi(\sigma\,|\,\lambda) \propto \sigma^{-1}$, so that the appropriate operational prior is

$$\pi_\lambda(\lambda,\sigma) = \pi(\lambda)\,\pi(\sigma\,|\,\lambda) \propto (1+\tfrac{1}{2}\lambda^2)^{-\frac{1}{2}}\,\sigma^{-1}$$

or, in terms of the original metric,

$$\pi_\lambda(\mu,\sigma) = (1+\tfrac{1}{2}\lambda^2)^{-\frac{1}{2}}\,\sigma^{-2}.$$

The corresponding reference posterior density of λ is

$$\pi(\lambda\,|\,z) = \pi(\lambda\,|\,r)$$
$$\propto (1+\tfrac{1}{2}\lambda^2)^{-\frac{1}{2}}\left\{\exp(-\tfrac{1}{2}n\lambda^2)\int_0^\infty \omega^{n-1}\exp(-\tfrac{1}{2}\omega^2+r\lambda\omega)\,d\omega\right\}.$$

One may observe that the factor in brackets is proportional to $p(r\,|\,\lambda)$ and thus the marginalization paradox does not occur. Similar results are obtained with the other examples in Dawid *et al.* (1973). We conjecture that our procedure always avoids the marginalization paradoxes; however, we do not have a proof.

5.2. *The Fieller–Creasy Problem*

In biological assay work one is often interested in the relative power of two treatments on drugs, and the following problem suggests itself. Suppose that two samples $\mathbf{x} = \{x_1, \ldots, x_m\}$ and $\mathbf{y} = \{y_1, \ldots, y_n\}$ are available from two independent normal populations with unknown means μ, η and common unknown variance σ^2. The problem is to make inferences about the value of $\theta = \mu/\eta$, the ratio of the means.

This problem was discussed in a symposium on Interval Estimation held by this Society. Fieller (1959) and Creasy (1959) presented there two different solutions that both claimed to be fiducial. Fieller's solution, defended by R. A. Fisher in the discussion, is difficult to accept for it can lead, for instance, to a "confidence" interval consisting of the *whole* real line. Kappenman *et al.* (1970) showed that Creasy's solution may be reproduced from a Bayesian point of view by the use of the familiar "non-informative" prior $\pi(\mu,\eta,\sigma) \propto \sigma^{-1}$.

In a previous paper (Bernardo, 1977a) we obtained the reference prior to make inferences about $\theta = \mu/\eta$ using the procedure described above. In terms of $\{\theta,\eta,\sigma\}$ such a prior turns out to be

$$\pi_\theta(\theta,\eta,\sigma) = \pi(\theta)\,\pi(\eta\,|\,\theta)\,\pi(\sigma\,|\,\eta,\theta) \propto (1+\theta^2)^{-\frac{1}{2}}\,\sigma^{-1}$$

or, in terms of the original parameters,

$$\pi_\theta(\mu, \eta, \sigma) \propto (\mu^2+\eta^2)^{-\frac{1}{2}}\,\sigma^{-1}.$$

The corresponding reference posterior distribution of θ after the samples $\mathbf{x}$ and $\mathbf{y}$ have been observed is

$$\pi(\theta\,|\,\mathbf{x},\mathbf{y}) \propto (1+\theta^2)^{-\frac{1}{2}}\left\{(n+\theta^2 m)^{-\frac{1}{2}}\left(S^2+\frac{mn(\bar{x}-\theta\bar{y})^2}{n+\theta^2 m}\right)^{-\frac{1}{2}(m+n-1)}\right\}, \tag{24}$$

where $\bar{x} = \Sigma x_i/m$, $\bar{y} = \Sigma y_i/n$, and $S^2 = \Sigma(x_i-\bar{x})^2+\Sigma(y_i-\bar{y})^2$. This is of the form

$$\pi(\theta\,|\,\mathbf{x},\mathbf{y}) \propto (1+\theta^2)^{-\frac{1}{2}}p(\mathbf{x},\mathbf{y}\,|\,\theta),$$

where the term in brackets in (24), $p(\mathbf{x},\mathbf{y}\,|\,\theta)$ is an integrated likelihood which, as one would expect, coincides with the integrated likelihood derived by Kalbfleish and Sprott (1970) for this example.

The reference posterior (24) has been studied using Monte Carlo methods with satisfactory results. Clearly, it is a symmetric density about the origin when either $\bar{x} = 0$ or $\bar{y} = 0$. This is to be expected since, in either case, there is no information to decide on the sign of θ. This feature is not obtained with the usual prior $\pi(\mu, \eta, \sigma) = \sigma^{-1}$.

5.3. *Stein's Paradox*

Marginalization paradoxes may be considered to be a powerful argument against the use of a unique reference prior for a given model. Since those paradoxes disappear when one uses proper priors, one is tempted to blame impropriety for the unsatisfactory results often obtained in multi-parameter situations with the usual improper operational priors. However, to use proper approximations to those priors when trying to describe the inferential content of the data does not work either. This is clearly demonstrated in Stein's (1959) example on the sum of the squares of normal means. Indeed, the universally recommended operational prior for a multivariate normal model with known precision matrix is $\pi(\mu_1, \ldots, \mu_k) \propto 1$, which we certainly regard as appropriate to produce reference posterior distributions for any set of the μ_i's, and this prior may be approximated by the proper density $p(\mu_1, \ldots, \mu_k) = \pi N(\mu_i\,|\,0, \sigma)$ where σ is very large. Now, suppose that we desire to make inferences about the value of $\theta = \Sigma\mu_i^2$; it is easily verified (Efron, 1973) that the use of such a prior overwhelms, for large k, what the data have to say about θ, so that the corresponding posterior distribution for θ is rather unsatisfactory.

From our point of view, the use of a uniform prior does not make sense if one is interested in θ; indeed, to obtain a reference posterior for θ we have to maximize the missing information about θ, a completely different situation to one in which you want to maximize the missing information about the μ_i's. We now turn to derive our reference posterior distribution for θ.

Let $\varepsilon(n)$ be the experiment which consists of n observations from each one of k independent normal distributions with means μ_i $(i = 1, \ldots, k)$ and variance 1. Let $\bar{x}_i$ be the mean of the n observations from population i, and let $\boldsymbol{\mu}$ and $\bar{\mathbf{x}}$ be the corresponding vectors in R^k. Thus, $p(\bar{\mathbf{x}}\,|\,\mu) = N(\bar{\mathbf{x}}\,|\,\boldsymbol{\mu}, n^{-1}\mathbf{I}_k)$ and $p(\boldsymbol{\mu}\,|\,\bar{\mathbf{x}}) \propto p(\bar{\mathbf{x}}\,|\,\boldsymbol{\mu})p(\boldsymbol{\mu})$. For large n, the prior density $p(\boldsymbol{\mu})$ may be ignored so that the asymptotic posterior distribution of $\boldsymbol{\mu}$ is $p^*(\boldsymbol{\mu}\,|\,\bar{\mathbf{x}}) = N(\boldsymbol{\mu}\,|\,\bar{\mathbf{x}}, n^{-1}\mathbf{I}_k)$ and therefore (see, for example, Graybill, 1961, chapter 4) with $\theta = \Sigma\mu_i^2 = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$ and $t = \Sigma\bar{x}_i^2 = \bar{\mathbf{x}}^{\mathrm{T}}\bar{\mathbf{x}}$, $n\theta$ has asymptotically a non-central χ^2 distribution with k degrees of freedom and parameter nt. It follows (see, for example, Johnson and Kotz, 1970, p. 139) that the posterior distribution of θ is asymptotically normal with variance $(2/n)\{2t+(k/n)\}$. Moreover, the sampling distribution of nt is a non-central χ^2 distribution with k degrees of freedom and parameter $n\theta$, so that $E(t\,|\,\theta) = \theta+(k/n)$ and therefore t is an asymptotically consistent estimate $\tilde{\theta}$ of θ. It now follows from the last paragraph of Section 3.3 that the reference prior for θ is $\pi(\theta) \propto \theta^{-\frac{1}{2}}$. One may note in passing that this could theoretically have been obtained from the sampling distribution of t, $p(t \mid \theta)$, assuming t conditionally sufficient, by the use of Jeffreys' formula; this proves to be however a difficult exercise in calculus.

Thus, if the conditional prior $p(\boldsymbol{\mu} \mid \theta)$ is such that t is sufficient or, alternatively, if inferences about θ are desired solely based on the value of t, the appropriate reference posterior is

$$\pi(\theta \mid t) \propto \pi(\theta) p(t \mid \theta) = \theta^{-\frac{1}{2}} \chi^2(nt \mid k, n\theta). \tag{25}$$

A student of mine, J. R. Ferrandiz, has recently shown that the same reference posterior is obtained without the assumption of sufficiency; thus, if one works in polar coordinates, in terms of θ and the corresponding vector $\boldsymbol{\omega}$ of angles, the reference prior to make inferences about θ is, for some function $f(\boldsymbol{\omega})$,

$$\pi_\theta(\theta, \boldsymbol{\omega}) = \pi(\theta)\pi(\boldsymbol{\omega} \mid \theta) \propto \theta^{-\frac{1}{2}} f(\boldsymbol{\omega})$$

and the corresponding reference posterior distribution for θ is again (25).

In his recent address to this Society, Wilkinson (1977) makes Stein's example central for his argument of "fiducial" versus Bayesian inference. We proceed to compare his solution with ours. Indeed, with the data he uses, i.e. with $n = 1$, $k = 50$ and $t = 100$, the 95 per cent shortest credible interval for θ is (19·4, 88·2) as derived from (25) by numerical integration. This is not far from the fiducial interval (21, 89) which he quotes.

Consider however the data $n = 1$, $k = 10$ and $t = 9{\cdot}133$. The value 9·133 for $t = \Sigma x_i^2$ was obtained by simulation as the sum of the squares of ten normal deviates with zero mean and unit variance. Thus, the "true" value of θ is 0. Note that there is nothing special about this value, since $p(t \mid \theta = 0)$ is a central χ^2 distribution with 10 degrees of freedom so that the value of t would be expected to lie between 6·7 and 12·5 with probability $\frac{1}{2}$. The corresponding posterior density of θ obtained from (25) decreases monotonically from 0 and, in particular, $P(\theta < 1 = t) = 0{\cdot}3952$ and $P(\theta < 5 \mid t) = 0{\cdot}7903$. The corresponding upper bounds obtained using Wilkinson's method are 0·6003 and 0·8247 but this leaves an "unassigned" probability of $p_0 = P\{\chi^2(10) > 9{\cdot}133\} = 0{\cdot}5195$ so that, for him, $p(\theta < 1 \mid t)$ could lie anywhere between 0·0808 and 0·6003. Wilkinson claims that "a high value of p_0 would indicate evidence that the observed point is too close to 0 to be statistically compatible with the assumed covariance matrix $\mathbf{I}_k$ of $\mathbf{x}$ or else with the normal form of the distribution". However, our data were obtained by simulation precisely from a multinormal distribution with $\mathbf{I}_k$ as covariance matrix!

Finally, as Smith (1977) clearly shows, Wilkinson's results are inconsistent with those directly obtained for the one-dimensional normal case. Indeed, using Smith's example, if one obtains $x = 1{\cdot}1503$ as a realization of a normal random variable with unknown mean μ and unit variance, the reference posterior distribution for μ is $\pi(\mu \mid x) = N(\mu \mid x, 1)$ so that $P(-1 < \mu < 1 \mid x = 1{\cdot}1503) = \int_{-1}^{1} N(\mu \mid 1{\cdot}1503, 1)\, d\mu = 0{\cdot}4245$. This is consistent with the result $P(\mu^2 < 1 \mid t = 1{\cdot}1503^2) = 0{\cdot}4245$ obtained using (25) with $n = k = 1$ and $t = x^2$, and one may prove that this is true for all x. This was to be expected since we have calculated in two alternative ways the probability of the same event, given the same information. This is to be compared with Wilkinson's rather surprising results $0{\cdot}2060 < P(\mu^2 < 1 \mid t = 1{\cdot}1503^2) < 0{\cdot}4560$, but $P(-1 < \mu < 1 \mid x = 1{\cdot}1503) = 0{\cdot}4245$ exactly!

6. DISCUSSION

The derivation of reference posterior distributions may be seen as a part of an analysis of the sensitivity of the posterior distribution to changes in the prior. The reference posterior distribution provides an origin for those statements about the parameter of interest which may be regarded as admissible, given the model and the data. Being an origin for admissible inferences, the reference posterior distribution need not be itself admissible but only arbitrarily close to admissible posteriors; indeed zero, which is not positive, is an appropriate origin for positive quantities.

In a private conversation, G. A. Barnard suggested to me the appealing name *information-modulated likelihood function* for the product $\pi(\theta)p(x|\theta)$. However, no claim of "objectivity" is made for the set of inferences that could be produced from its normalized form, the reference posterior $\pi(\theta|x)$. It is only argued that $\pi(\theta|x)$ gives a coherent feeling of the values of θ that the data x are supporting, under the assumptions that $p(\theta)\in C$ and that the model is true. One should compare the reference posterior $\pi(\theta|x)$ with the posterior density $p(\theta|x)$ obtained from a personal prior $p(\theta)$ which describes the scientist's initial information; the distance between $p(\theta|x)$ and $\pi(\theta|x)$ would be a measure of the relevant information contained in $p(\theta)$.

A reference prior does *not* describe a situation of "non-information" about the parameters of a model; the examples in Section 5 show that such a description is not possible. Instead, if (θ,ω) are the parameters of the model, $\pi_\theta(\theta,\omega)$ describes the limit of a particular kind of knowledge about (θ,ω): that which leaves most to be learned from the data about the value of θ. This is why, although invariant to one-to-one transformation of the parameter space, the method is *not* invariant to marginalization. We maintain that the reference posterior which corresponds to such a prior is a useful distribution to quote in a scientific report about θ.

If it is desired to restrict the sensitivity analysis to some specific class C of priors, e.g. those compatible with some accepted information or those introducing some assumptions, this is done by maximizing the missing information in C rather than in the class of all probability densities. Although in this paper we have only worked in the latter case, we believe this is a promising field of research. It could be used, for instance, to derive reference priors for the last step of hyperparameters in a hierarchical prior specification as those used by Lindley and Smith (1972) and Smith (1973); here, C would be the class of priors with the assumed hierarchical structure, and one would have to find a reference prior by maximizing in C the missing information about the parameter of interest.

It should be clear to the audience that an entirely satisfactory mathematical presentation of the methods suggested in this paper would require much more attention to detail than has been attempted here. In particular, the asymptotic behaviour of posterior entropies, and the maximization process which Definition 1 requires, should be more carefully investigated. However, an understanding for the foundations and consequences of the procedure advocated here can be achieved with the informal approach adopted.

I would like to conclude by quoting the last paragraph of Professor Novick's address to this Society (Novick, 1969) on precisely the same topic I have been discussing tonight, for it describes precisely my own feelings: "The paper is put forward as a further foray into the unknown to see if the basic principles are sound. Naturally, the emphasis has been on the case for the defence, though no relevant, possibly embarrasing facts have been suppressed. The case for the prosecution will, I am sure, follow shortly".

ACKNOWLEDGEMENTS

The author is very grateful to Professors G. A. Barnard, J. A. Hartigan, D. V. Lindley, I. R. Savage, A. F. M. Smith and C. Villegas and to the referees by their detailed, encouraging comments on earlier drafts of this paper.

REFERENCES

AKAIKE, H. (1978). A new look at the Bayes procedure. *Biometrika*, **65**, 53–59.

BARNARD, G. A. (1952). The frequency justification of certain sequential tests. *Biometrika*, **39**, 144–150.

BARTHOLOMEW, D. J. (1965). A comparison of some Bayesian and frequentist inferences. *Biometrika*, **52**, 19–36.

BAYES, T. R. (1763). Essay towards solving a problem in the doctrine of changes. Reprinted in *Biometrika*, **45** (1958), 243–315.

BERNARDO, J. M. (1975). Non-informative prior distributions: a subjectivist approach. *Bull. Internat. Statist. Inst.*, **46**, 94–97.

—— (1977a). Inferences about the ratio of normal means: a Bayesian approach to the Fieller–Creasy problem. In *Recent Developments in Statistics* (J. R. Barra *et al.*, eds), pp. 345–349. Amsterdam: North-Holland.

Bernardo, J. M. (1977b). Inferencia Bayesiana sobre el coeficiente de variación: una solución a la paradoja de marginalización. *Trab. Estadist.*, **28**, 23–30.

—— (1978). Una medida de la información útil proporcionada por un experimento. *Rev. Acad. Ci. Madrid*, **72**, 419–440.

—— (1979). Expected information as expected utility. *Ann. Statist.*, **7** (to appear).

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass.: Addison-Wesley.

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

Creasy, M. A. (1959). Limits for the ratio of the means. *J. R. Statist. Soc.* B, **16**, 186–199.

Dawid, A. P. (1970). On the limiting normality of posterior distributions. *Proc. Camb. Phil. Soc.*, **67**, 625–633.

Dawid, A. P., Stone, N. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. R. Statist. Soc.* B, **35**, 189–233 (with discussion).

DeGroot, M. H. (1970). *Optimal Statistical Decision.* New York: McGraw-Hill.

Dickey, J. M. (1973). Scientific reporting and personal probabilities: Student hypothesis. *J. R. Statist. Soc.* B, **35**, 285–305.

—— (1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.*, **71**, 680–689.

Efron, B. (1973). In discussion of Dawid, Stone and Zidek (1973). *J. R. Statist. Soc.* B, **35**, 219.

Fieller, E. C. (1954). Some problems in interval estimation. *J. R. Statist. Soc.* B, **16**, 186–194 (with discussion).

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. R. Statist. Soc.* B, **25**, 368–376.

Good, I. J. (1966). A derivation of the probabilistic explanation of information. *J. R. Statist. Soc.* B, **28**, 578–581.

—— (1969). What is the use of a distribution? *Multivariate Analysis* (Krishnaiah, ed.), Vol. II, pp. 183–203. New York: Academic Press.

Graybill, F. A. (1961). *An Introduction to Linear Statistical Models.* New York: McGraw-Hill.

Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika*, **35**, 297–303.

Hartigan, J. A. (1964). Invariant prior distributions. *Ann. Math. Statist.*, **35**, 836–845.

—— (1965). The asymptotically unbiased prior distribution. *Ann. Math. Statist.*, **36**, 1137–1152.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics*, **SCC-4**, 227–291.

—— (1978). Marginalization and prior probabilities. To appear in *Studies of Bayesian Statistics* (A. Zellner, ed.). Amsterdam: North-Holland.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. London* A, **186**, 453–461.

—— (1939/67). *Theory of Probability* (3rd ed.). Oxford: Clarendon Press.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions.* Boston: Houghton Mifflin.

Johnston, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.*, **41**, 851–854.

Kalbfleish, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large number of parameters. *J. R. Statist. Soc.* B, **32**, 175–290.

Kappenman, R. F., Geisser, S. and Antle, C. E. (1970). Bayesian and fiducial solutions to the Fieller–Creasy problem. *Sankhyā* B, **32**, 331–340.

Kashyap, R. L. (1971). Prior probability and uncertainty. *IEEE Trans. Information Theory* **1T-14**, 641–650.

Laplace, P. S. (1825). *Théorie des Probabilités.* Reprinted (1960). Paris: Courcier.

Lee, P. M. (1964). On the axioms of information theory. *Ann. Math. Statist.*, **35**, 415–418.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.*, **27**, 986–1005.

—— (1958). Fiducial distributions and Bayes theorem. *J. R. Statist. Soc.* B, **20**, 102–107.

—— (1961). The use of prior probability distributions in statistical inference and decisions. *Proc. 4th Berkeley Symp.*, **1**, 436–468. Berkeley: University of California Press.

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

—— (1971). *Bayesian Statistics, a review.* Reg. Conf. Ser. Appl. Math., **2**. Philadelphia: SIAM.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Statist. Soc.* B, **34**, 1–42.

Novick, M. R. (1969). Multiparameter Bayesian indifference procedures. *J. R. Statist. Soc.* B, **31**, 29–64 (with discussion).

Novick, M. R. and Hall, W. J. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.*, **60**, 1104–1117.

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries*, **73**, 285–334.

Piccinato, L. (1973). Un metodo per determinare distribuzioni iniziali relativamente non-informative. *Metron*, **31**, 1–13.

Piccinato, L. (1978). Predictive distributions and non-informative priors. *Trans. 7th Prague Conf. Information Theory* (in press).

Rényi, A. (1964). On the amount of information concerning an unknown parameter in a sequence of observations. *Publ. Math. Inst. Hung. Acad. Sci.*, **9A**, 617–625. Reprinted in *Selected papers of Alfred Rényi* (Turán, ed.) (1976), pp. 272–279. Akadémiai Kiadó.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, **27**, 379–423, 623–656.

Smith, A. F. M. (1973). A general Bayesian linear model. *J. R. Statist. Soc.* B, **35**, 67–75.

—— (1977). In discussion of Wilkinson (1977). *J. R. Statist. Soc.* B, **39**, 145–147.

Stein, C. (1956). Inadmissibility of the usual estimation for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp.*, (J. Neyman and E. L. Scott, eds), **1**, 197–206. Berkeley: University of California Press.

—— (1959). An example of wide discrepancy between fiducial and confidence interval. *Ann. Math. Statist.*, **30**, 877–880.

Stone, M. (1965). Right Haar measures for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.*, **36**, 440–453.

—— (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.*, **41**, 1939–1953.

—— (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.*, **71**, 119–125 (with discussion).

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayesian inference in routine statistical problems. *Biometrika*, **59**, 269–375.

Villegas, C. (1971). On Haar priors. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 409–414. Toronto: Holt, Rinehart & Winston.

—— (1977a). Inner statistical inference. *J. Amer. Statist. Assoc.*, **72**, 453–458.

—— (1977b). On the representation of ignorance. *J. Amer. Statist. Assoc.*, **72**, 651–654.

Walker, D. M. (1969). On the asymptotic behaviour of a posterior distribution. *J. R. Statist. Soc.* B, **31**, 80–88.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *J. R. Statist. Soc.* B, **25**, 318–329.

Wilkinson, G. N. (1977). On resolving the controversy in statistical inference. *J. R. Statist. Soc.* B, **39**, 119–171 (with discussion).

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

—— (1977). Maximal data information prior distributions. *New Developments in the Applications of Bayesian Methods* (A. Aykac and C. Brumat, eds), pp. 211–132. Amsterdam: North-Holland.

Discussion of Professor Bernardo's Paper

Professor J. B. Copas (University of Salford): I would like to start by welcoming Professor Bernardo to the Society and complimenting him on the presentation of his case both tonight and in his written paper. I believe we have before us a paper which is both important and challenging, and I am pleased to have the privilege of opening the discussion. There are several points I would like to take up, but perhaps it is incumbent on the opening prosecution witness to start off by taking his spade to the very roots of the edifice, leaving it to the later witnesses, many more expert than I, to comment on more particular matters.

The backbone of the method is equation (1), Professor Bernardo's idea being that to maximize this expression one maximizes the contribution of the data and minimizes the contribution of the prior distribution. The argument rests on the belief that the entropy of a distribution is a measure of uncertainty. Consider the distribution shown in Fig. D1. If this is cut in half, and the two halves moved apart, the variance increases dramatically. For example, if we are forecasting next year's company profits, then the original distribution says that we are sure to break almost even over the year, whereas the displaced distribution says we are sure to make almost £1m profit or £1m loss, but we have no idea which. Surely, the company is operating under much greater uncertainty in the second case than in the first, yet the entropies of the two distributions are exactly the same. This is because entropy depends only on the distribution of the different heights of the probability function, and pays no regard to the values of the variable at which these various heights are attained. Entropy is the average amount of information which has to be transmitted in order to specify without error which particular value of a random variable is obtaining at any particular time, a quite different matter from measuring the statistical uncertainty in the value of the random variable itself. Given, then, that entropy is a very imperfect measure of statistical uncertainty, how does Professor

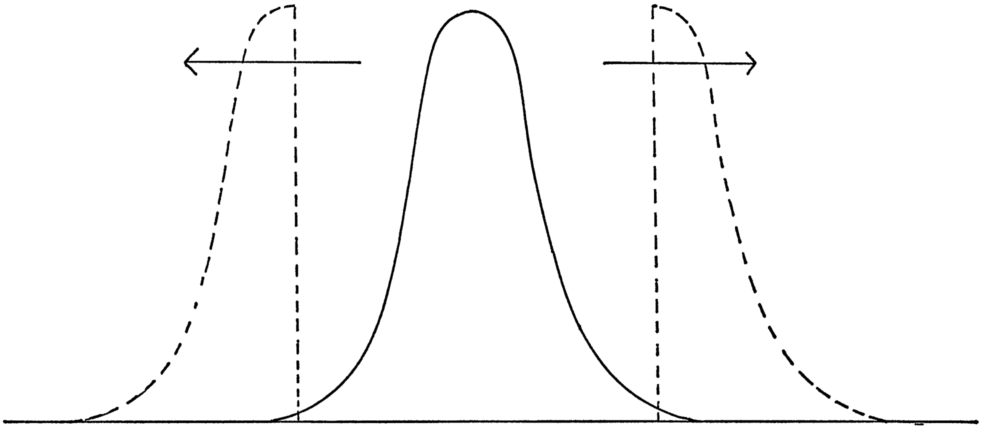

FIG. D1.

Bernardo's method apparently remain unscathed? It is because asymptotic posterior distributions are usually normal, when the entropy is essentially the log of the standard deviation and thus a monotonic function of variance.

If the asymptotic posterior of θ is independent of the prior, then the equation leading to (3) can be written

$$M(p(\theta)) - E_\theta\, \alpha(\theta), \tag{A}$$

where

$$\alpha(\theta) = E\{M(p(\theta \mid z)) \mid \theta\}$$

and M is the measure of uncertainty, taken in the paper to be entropy. Here, once the experiment is specified, the function $\alpha(\theta)$ is fixed, and expression (A) is the quantity to be maximized over $p(\theta)$. To minimize the second term, the prior should concentrate on those values of θ which make $\alpha(\theta)$ small, but on the other hand, to maximize the first term requires probability over a wide range of values; Professor Bernardo's solution is the compromise between these two opposing forces, and naturally the resulting prior density in (5) is a monotonically decreasing function of $\alpha(\theta)$. An important special case is when $\alpha(\theta)$ is constant, as for instance happens for an unknown location parameter. One is then left to maximize just $M(p(\theta))$, which by any reasonable definition of M will spread the probability out to a uniform distribution. Thus, as Professor Bernardo rightly says, the uniform prior does not require asymptotic normality, but neither does it require the use of entropy; it would just as well result if M was variance, assuming one only wished to optimize within unimodal distributions. Similarly, the uniform prior on log θ would result for a scale parameter.

If (A) is univariant under one-to-one transformations of θ, then this argument extends to the regular continuous case. For in the notation of the paper, the asymptotic posterior distribution of the transformed parameter ϕ defined by

$$\phi = \int^\theta i(\theta)^{\frac{1}{2}}\, d\theta$$

is normal with variance independent of ϕ, and so ϕ is assigned a uniform prior distribution, or θ itself the Jeffreys' prior as in (13). The essential property required of M is that (A) be invariant, not specifically, that it be entropy. It would be interesting to know whether there is some other measure M which more directly relates to statisticians' ideas of uncertainty and yet which leaves (A) invariant. If such a measure exists, it might form a better rationale for the results derived in tonight's paper.

As I have remarked already, entropy pays no regard to the metric of the sample space of the relevant random variable, and so can take no account of smoothness of the resulting distribution. I think this is another difficulty with entropy. Professor Bernardo emphasizes that his method can apply to the situation when one wants to incorporate some specific knowledge about θ by maximizing within the restricted class C of prior distributions consistent with that knowledge. Perhaps this is the most promising aspect of his technique but unfortunately the results can be somewhat unappealing. For instance, suppose we wish to incorporate the knowledge that the prior probability of θ belonging to some set S is p. Then a straightforward extension of the analysis given in Section 3 shows that in the regular continuous case the prior distribution for θ is

$$\pi(\theta) = \begin{cases} c_1\, i(\theta)^{\frac{1}{2}}, & \theta \in S, \\ c_2\, i(\theta)^{\frac{1}{2}}, & \theta \notin S, \end{cases}$$

where c_1 and c_2 are chosen such that $P(\theta \in S)$ and $P(\theta \notin S)$ are proportional to p and $(1-p)$ respectively. The discontinuities on the boundaries of S do not make much sense from a practical point of view, since the choice of both S and p are likely to be somewhat arbitrary. However, when C takes the form of specifying prior *moments*, Professor Bernardo's technique can give a simple and rather appealing solution. For instance, if the prior mean and variance of θ are to be fixed in advance, one obtains the solution

$$\pi(\theta) \propto i(\theta)^{\frac{1}{2}} \exp(\lambda_1\, \theta + \lambda_2\, \theta^2),$$

where λ_1 and λ_2 are specified in order to give $\pi(\theta)$ the required mean and variance. The first term is the distribution obtained when no information about θ is assumed, and the second term is simply a normal density. Interestingly, this is roughly equivalent to assuming one has available the data from a supplementary sampling experiment which gives rise to a "normal" likelihood function which is then multiplied in accordance to Bayes' theorem.

Professor Bernardo points out that his method is not invariant under many-to-one transformations or under marginalization. As he says, this means that one's choice of prior depends on which aspect of the parameter is under study. But there are cases where the choice of a "natural" parameterization is not clear. For instance, in the counterfeit coin example in Section 4.1, suppose we are interested in whether the coin is double-headed (ψ_1). What is the alternative hypothesis? If it is the composite of ψ_0 and ψ_2, then $P(\psi_1) = \frac{1}{2}$, but if there are two separate simple alternatives ψ_0 and ψ_2, $P(\psi_1) = \frac{1}{3}$. Does the author's analysis of this example imply that the probability of *any* simple hypothesis that we care to examine in the finite discrete case is $\frac{1}{2}$? I find the discussion of this in the paper less than adequate. Similarly, in Stein's paradox in Section 5.3 one may be interested in making separate decisions for each component problem. As such, the complete vector of parameters $\boldsymbol{\mu}$ is the object of inference. But it so happens that, using a combined loss function, risks of symmetric decision rules depend on $\boldsymbol{\mu}$ only through a scalar function such as $\theta = \Sigma\, \mu_i^2$. Is then θ the object of inference? More generally, the parameter in a decision problem may not be an object of inference at all, but simply that part of the model which links the loss function to the likelihood function.

If tonight's speaker is serious in his claim that the method is consistent with a subjectivist view of probability, then reference prior distributions cannot possibly be interpreted as inferences in their own right. In the tone of his discussion of the examples in Sections 4 and 5, however, I detect that Professor Bernardo comes very close to interpreting them as if they in fact are. One of the most compelling consequences of the Bayesian argument is that inferences can be updated in a sequential way as new information arises; this too cannot be so for reference posterior distributions, as the complete form of the likelihood function has to be known before the initial prior can be formulated. Perhaps Professor Bernardo is himself near the brink of his own trap, but I think there is a great danger that some, who from indoctrination believe they should always find Bayesian solutions to data problems, will dive into the trap headlong and interpret the method of tonight's paper as a recipe for prescribing prior distributions which do indeed represent ignorance. One is back to the position of interpreting the reference posterior merely as a yardstick. But what is the use of a yardstick if we do not know how to measure with it? I look forward to hearing Professor Bernardo expand on his meaning of "origin" and "reference".

Tonight's paper has been stimulating and provocative, but as perhaps should always be the case with a good read paper, there are many questions left unanswered. It gives me great pleasure to propose the vote of thanks.

Dr A. O'HAGAN (University of Warwick): "Ignorance is bliss", they say, but the question of whether it really serves any useful purpose to represent prior ignorance formally is highly contentious. Nevertheless, I will confine my remarks to operational behaviour of Professor Bernardo's reference posteriors, because despite his protestations that they only represent an "origin" or "reference" it is clear from his examples that he sees them as being meaningful, and perhaps useful, in themselves.

It is obviously a significant achievement to derive, from a single framework, the uniform prior for "finite discrete" cases and the Jeffreys' prior for suitable continuous cases. Moreover, it is very important that Professor Bernardo has indicated regularity conditions for the Jeffreys' prior to be appropriate. Consider, for instance, a simple class of problems where the posterior distribution is not asymptotically normal and therefore Jeffreys' prior is not obtained: let the observation x have density function given θ of the form

$$p(x \mid \theta) = f(x, \phi(\theta)),$$

depending on θ only through the function $\phi(\theta)$. Then if $\phi(\theta)$ is not a one-to-one transform, the whole parameter θ is not identified. The Jeffreys' prior is clearly a function of $\phi(\theta)$ alone, implying that the conditional prior distribution of θ given $\phi(\theta)$ is uniform. Now it is well known that in such a case the data do not modify the distribution of θ given $\phi(\theta)$, so it will be uniform in the posterior also, which may well result in the posterior being improper. But Professor Bernardo's basic approach of Section 2 recognizes the fact that the data tell us nothing about θ given $\phi(\theta)$ and, quite properly, admits defeat—the reference prior for θ given $\phi(\theta)$ turns out to be arbitrary. This in turn highlights the fact that posterior distributions, reference or otherwise, require a conscious specification of prior information.

But whereas I am impressed by Section 2, I find Professor Bernardo's handling of nuisance parameters in Section 4 much less convincing. We are treated to some elegant verbal sidestepping but the basic idea must still amount to a way of representing total ignorance about the parameter $\psi = (\theta, \omega)$. First we pretend ignorance about ω for every possible value of θ, yielding a set of conditional reference priors $\pi(\omega \mid \theta)$. Then we pretend ignorance about θ to obtain a reference prior $\pi(\theta)$. Yet the result of multiplying these two is different from pretending ignorance about ψ directly. To see how different they can be, consider the "counterfeit coin" example of Subsection 4.1. The direct approach yields what I will call the unconditional reference prior $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ with entropy log 3. Using the methods of Section 4 to obtain $\pi(\omega \mid \theta)\, \pi(\theta)$ yields what I will call the conditional reference prior $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\}$ with entropy 1·5 log 2, which is only 5 per cent less than log 3. This is only a small difference, but suppose we extend the example to allow the coin to be biased not just to the two extremes of double-headed or double-tailed but to k different degrees. The unconditional and conditional reference priors are

$$\left\{\frac{1}{k+1}, \ldots, \frac{1}{k+1}\right\} \quad \text{and} \quad \left\{\frac{1}{2}, \frac{1}{2k}, \ldots, \frac{1}{2k}\right\}$$

respectively, and the entropy of the latter for large k is only about half that of the former. So the conditional reference prior contains up to half of the missing information: how can these both be representations of total ignorance? Even greater discrepancies can be achieved with multinomial sampling. Professor Bernardo tries to justify his conditional reference prior by arguing that a single toss of a coin tells us nothing about whether it is fair, but the Bayesian argument acknowledges this fact regardless of the prior—by the prior and posterior probabilities that the coin is fair being equal.

But even if the conditional priors are sensible, by changing which aspect of the parameter we regard as being a nuisance we change the reference prior, and hence the reference posterior, for the full parameter. In his very first sentence Professor Bernardo invokes coherence to justify being a Bayesian, and yet his approach to nuisance parameters is incoherent. Referring again to the counterfeit coin, imagine that a single toss of the coin results in heads. Then the reference posterior distribution for inference about whether it is fair is $\{\frac{1}{2}, \frac{1}{2}, 0\}$. But the reference posterior for inference about whether it is double-headed is $\{\frac{1}{5}, \frac{4}{5}, 0\}$. If asked to bet on whether the coin is fair, Professor Bernardo refers to the first posterior distribution and will accept any odds better than evens. And if asked to bet on whether it is double-headed he refers to the second distribution and accepts odds better than 4–1 on. It would be very easy, with these highly incoherent beliefs, for him to place bets which would lose him money whatever the true state of the coin, and yet which he would believe were to his advantage! I would like to be his bookmaker!

So inferences about different aspects of the full parameter do not cohere, and the prime symptom is that probability laws fail—the same probability evaluated two different ways yields two different answers. As another example consider evaluating the probability that the next toss of the coin will also be heads. We should completely reformulate the problem so as to express the result of the next toss as a function of the parameter (and I would be interested to see how Professor Bernardo would do this), but one might be tempted, naively, to use probability calculus via

$$P(\text{heads next}) = P(\text{heads next} \mid \text{fair})\, P(\text{fair}) + P(\text{heads next} \mid \text{double-headed})\, P(\text{double-headed}).$$

With the above reference posteriors we find

$$P(\text{heads next}) = \tfrac{1}{2} \times \tfrac{1}{2} + 1 \times \tfrac{4}{5} = 1{\cdot}05!$$

Although I think Professor Bernardo would have been safer sticking to his unconditional theory of Section 2, I am glad that he did not. As a result he has given us a paper which is not only lucid and stimulating but also challenging. It gives me very great pleasure to second the vote of thanks.

The vote of thanks was carried by acclamation.

Professor A. F. M. SMITH (University of Nottingham): It has long been a source of considerable embarrassment to dwellers in the Bayesian citadel that it houses so many improper waifs and strays, most of hideously deformed appearance. How fitting that a Dr Bernardo should appear with the aim of providing a respectable shelter for these outcasts!

My own view of "vague" or "improper" priors is that they are simply mathematical artefacts (having no intrinsic interest in their own right) whose justification rests on the fact that their use in Bayes' theorem results in a posterior distribution which is a "good approximation", in some sense, to what would have been obtained using the "non-informative" prior anticipated from careful assessment. It is clear that the quality of an approximation will depend both on the parameter of interest and on the likelihood, and so it should not be a matter of surprise, or concern, if the form of representation of the "vague" prior (i.e. mathematical artefact) depends on the data, or does not transform in an "obvious" way when we change the parameter of interest. I certainly have no objection to Bernardo's results on these grounds.

But, from this "pragmatic" standpoint, how should one react in general to Professor Bernardo's rather formal approach? In a sense, if we take the "good approximation" idea seriously, then the whole business seems rather circular. An "actual" prior is "non-informative" (by definition!) only if the posterior it would lead to is well approximated by Bernardo's reference posterior. A possible alternative reaction is to note that the reference recipes seem intuitively satisfying and also provide an elegant unification and clarification of many hitherto messy issues. I suggest that we should, therefore, be pragmatically delighted with this paper, whilst continuing to bear in mind that *approximation* is the real issue.

I have some queries. First, there seems to be a promise, in the Introduction, to shed further light on the Stein inadmissibility result (Stein, 1956) and the strong inconsistency result of Stone (1976), in so far as they relate to particular improper prior representations. This promise does not appear to be fulfilled. Secondly, have some "possibly embarrassing facts" been inadvertently suppressed following equation (18)? The author has noted that, in the presence of nuisance parameters, the reference prior for θ will generally depend on $p(\omega \mid \theta)$. One option would appear to be to use the reference form $\pi(\omega \mid \theta)$. But is this an unambiguous procedure if $\omega = (\omega_1, \omega_2)$, say? We could proceed directly to obtain $\pi(\omega_1, \omega_2 \mid \theta)$, or we could obtain $\pi(\omega_1 \mid \theta, \omega_2)$, $\pi(\omega_2 \mid \theta)$ in two stages. Would the author comment on whether these alternatives necessarily lead to the same results? And what should we do in cases where they do not?

Finally, I should like to ask the author whether he feels his approach can help with the following important class of problems. Suppose we have a finite list of alternative models (for example, location-scale families with different tail behaviours, or alternative regression models) and wish to obtain posterior probabilities on the individual models, having assigned "non-informative" priors to parameters within each model. Should the non-informative priors for location and scale differ from family to family? If so, how? And what are the appropriate "constants" for "uniform" priors assigned to alternative vectors of regression coefficients having different dimensionalities?

Professor A. P. DAWID (The City University): I feel well placed to appreciate Professor Bernardo's achievement, since some time ago I myself tried, and failed, to carry out a similar programme. My approach was to consider an *uncertainty function* $U(\Pi)$ defined for distributions π for the parameter. This might, as in tonight's paper, be the entropy of Π, but I was thinking in terms of a decision problem with specified loss function, and taking U to be the expected loss consequent on taking the optimal decision for the state of information Π. Defining Π_0 as the prior distribution for θ, and Π_x the posterior based on data $X = x$, the expected value of sample information in the experiment is $U(\Pi_0) - E[U(\Pi_x)]$. It seems reasonable that an "uninformative" prior (relative to U) is one for which this quantity is maximized.

This idea occurred to me and to several others at about the same time, but no simple general solution emerged. While some special cases may be solved, these give little insight. Moreover, some of these answers seemed somehow "wrong".

Professor Bernardo has hit on the idea of maximizing the expected value of information from a large number of replications of the experiment. This gives more elegant and more acceptable answers. I should like to know if the method might extend to a general uncertainty function. For one based on a decision problem, we would actually get a "reference decision" for any observation.

This raises a general problem of interpretation. Reference posteriors (or decisions) are not for use: they are for reference. But just how are we supposed to make the comparison between our real (informed) analysis and the reference? And what use are we to make of this comparison?

A further problem arises from the "incoherence" of reference priors for different parameters. Suppose, for example, $\Delta = \theta_2 - \theta_1$ represents the effect of applying some treatment, and we are interested in whether the treatment has a positive effect. We want a reference posterior probability $P(\Delta > 0)$. What reference prior is called for? One might use that appropriate for inference about Δ, and integrate the posterior reference density over the event "$\Delta > 0$". But one could also construct a new parameter: $\Phi = 1$ if $\Delta > 0$, $\Phi = 0$ otherwise; and use a reference prior for Φ to obtain a (presumably) different reference value for $P(\Phi = 1)$. In other words, if we want to find the reference posterior probability of a set in the parameter space, this is *not* done by integrating the reference posterior density over that set. But if this is so, what use *are* we to make of reference densities?

Professor D. J. Bartholomew (London School of Economics): For a long time now I have only been a spectator in the game of hunting the prior but that, perhaps, provides a vantage point from which to make observations and ask questions. The author is to be congratulated for providing what, to me at least, is a convincing and impressive method of deriving what used to be called ignorance priors. It gives me some satisfaction to note that in the regular continuous case the Jeffreys' prior to which the method leads is also the one which Welch and Peers (1963) arrived at using frequentist arguments. In this connection it is worth observing (though somewhat remote from the rarified atmosphere of this evening's meeting) that we now have a further justification for some of the inference procedures which form the backbone of the elementary courses which we teach. I must confess to some slight amusement at the verbal manoeuvres in which the author has had to engage to maintain his Bayesian faith but, even so, he comes perilously close to heresy. For example, reference priors involve integration over the sample space and hence they depend on the sampling rule. It would be interesting to know how much the form of the reference prior is affected by the choice of sampling rule. Would it make sense, I wonder, to use the authors' method with sequential sampling schemes where the stopping rule might depend on the observations obtained to date. If so the relationship between the sampling rule and the prior might throw some further light on whether the basic principles are sound. Another direction in which the author might extend his work is to predictive inference where we are interested in the values of future observations rather than in parameters. This is a field where Bayesian methods are attractive from a practical point of view. Does the author think that his information criterion provides an approach to the choice of prior for that problem? There was much discussion of the correct choice of prior in the case of observations on Bernoulli variables following Thatcher (1964).

Professor D. V. Lindley (Somerset): The author of today's paper is to be congratulated on the ingenious ways that he has overcome the difficulties usually associated with reference priors. However one snag remains: the distributions derived by his procedure violate the likelihood principle, and therefore violate the requirements of coherence that he mentions. It is easy to see this because the method depends on repetitions, not of the result of an experiment, but of results like those obtained in an experiment, namely those in the sample space. It is even more transparent in the regular case where the expectation operator is used (equation (11)).

An example is illuminating. Suppose r successes and $n-r$ failures have been observed in a Bernoulli sequence with unknown chance of success θ. If n is fixed (binomial sampling) the reference prior is proportional to $\{\theta(1-\theta)\}^{-\frac{1}{2}}$ (Section 3.4). In contrast, if r is fixed (Haldane sampling) this is replaced by $\{\theta^2(1-\theta)\}^{-\frac{1}{2}}$. Consequently if you or I were to make inferences about θ, we could use our personal probabilities, but if we wanted to engage in scientific reporting, then we would have to go back to the data $(r, n-r)$, which had so far proved adequate, and enquire what the sampling rule was. I find it unnatural that completely new information should be needed for scientific, as distinct from personal, reporting.

A further example sheds more light. The trinomial distribution with chances $\lambda\{1-(1-\delta)\,\theta\}$, $(1-\delta\lambda)\,\theta$ and $(1-\lambda)(1-\theta)$ occurs in the analysis of life-tables and yields numbers D of deaths, W of withdrawals and S of survivors in the three classes. Here we take δ to be a known value in $[0, 1]$. The likelihood factors into a function of θ times one of λ. It might be expected that the reference prior would factor, giving independence, but this appears not to be so. Suppose C is restricted to the class of priors that do factor. Then we can isolate λ say and have likelihood $\lambda^D(1-\delta\lambda)^S(1-\lambda)^W$. For $\delta = 0$ or 1 this is Bernoulli, yet the reference prior in neither case is $\{\lambda(1-\lambda)\}^{-\frac{1}{2}}$ suggested by Professor Bernardo. I find this strange. But he has been so successful in overcoming other difficulties, and the rewards of success would be so great, that I am sure he will be able to overcome these teasers.

Dr P. J. Brown (Imperial College, London): In the search for "non-informative" priors the subjective element is at least crucial in deciding between the benefits and drawbacks of each candidate prior. A difficulty for me with tonight's approach arises in the context of medical diagnosis presented here in a very simplified form. The essence of the problem has been touched on by both Professors Copas and A. F. M. Smith. In the decision theory framework the parameter of interest may not be fixed and the dimension of the parameter space may be changing.

In an important paper, Lindley (1978) in response to Hughes (1968), has considered the case of two multinomial populations described by two sets of probabilities θ_i for the first, ϕ_i for the second, $i = 1, \ldots, n$, $\Sigma\,\theta_i = 1$, $\Sigma\,\phi_i = 1$. Training data in the form of N observations from each population are available. The two populations may be thought of as two diseases and it is also envisaged that $n = 2^s$ so that the labels to the cells of the multinomial are considered as strings of s binary symptoms. A change from $n = 2^s$ to 2^{s+1} is equivalent to introducing an extra symptom. Suppose there is a single undiagnosed case and it is desired to predict its population of origin. Now one would hope that the sequence of prior distributions obtained by increasing the number of symptoms observed would be such that the expected diagnostic accuracy would be non-decreasing. Use of Jeffreys' multiparameter prior as suggested by Professor Bernardo for this regular continuous case (Dirichlet with indices $\frac{1}{2}$) for any two numbers of symptoms s, s' would mean that the priors would not cohere and as pointed out by Lindley this leads to problems such as those of Hughes (1968) where you can actually expect to do worse by observing extra symptoms. That you might expect to do no better is reasonable but to expect to do worse within a Bayesian framework is disturbing. In an as yet unpublished paper I have shown that in the simple case when the training sets are very large ($N \to \infty$) so that θ and ϕ are determined, the expected probability of correct classification p_n given by

$$p_n = \frac{1}{2} \sum_{i=1}^{n} E \max(\theta_i, \phi_i)$$

is monotone non-decreasing for coherent specifications of sequences of priors and may approach any limit in $[\frac{1}{2}, 1]$ as $n \to \infty$ (contrary to a conjecture of Lindley, 1978). However, there is no such guarantee of monotonicity if the priors are not coherent. Mr P. Rundell at Imperial College is working on priors which are coherent and also informative in this diagnostic situation.

That said, I found this a very interesting paper.

Professor C. A. B. Smith (University College London): A recently discovered missing page from *Alice in Statland* reads:

White Rabbit (to Alice): I've grown 10 lettuces in a magnetic field, and 10 unmagnetized, and weighed them. I want to know if magnetism has made them bigger. How can I find out?

Alice: Ask "Significant Statisticians Ltd: Enquiries" over there.

Significant Statisticians: (Looking at data.) We are in agreement: it depends on *why* you grow 20 lettuces. If you decided beforehand on the number 20, the magnetic field had a significant effect. If you stopped at 20 because you thought you had proved the point, the effect of magnetism was insignificant and presumably zero.

White Rabbit: I just put seed in on rainy days.

Significant Statisticians: What a way to proceed! Always consult a statistician before you do an experiment!

Alice: I don't see how intentions can influence the growth of the lettuces: these statisticians seem incoherent to me. Let's try the "Strict Savages Enquiry Office" over there. (Going in). Are you coherent?

Strict Savages: Of course. (Looking at the data.) Smith here calculates that magnetism increased the mean weight by 13 grams. Jones thinks it decreases it by 6 grams...

Alice: You're hardly united.

Strict Savages: We never are: we each rely on our own prior opinion.

Alice: Let's try the "United Bayesians" office over there. (Going in.) Are you coherent and united?

United Bayesians: Yes, of course. We decide beforehand on our shared prior beliefs using the "Fisher information" which depends on how you plan your experiment. Thus, if you decide you will weigh the lettuces exactly, we have one opinion. If you weigh them to the nearest 10 grams, we have another.

Stone and Dawid (walking in): Beware! Their views are paradoxical.

Alice: Well, let's try the Better Bernardians over there. (Entering enquiry office.) Are you coherent, united, and Stone-resistant?

Better Bernardians: Yes. But of course the answer depends on whether you're interested only in the mean effect of the magnetic field, which then comes to 16 grams, or whether you're interested in the variability as well, when the mean effect comes to only 12 grams.

Alice: But how do you work that out?

Better Bernardians: By taking the prior corresponding to the limit of posteriors maximizing the missing information...

Alice: What is "information"?

Better Bernardians: Negative expected log probability.

Alice: If negative expected log probability is information, then I am the Queen of Hearts. Off with all your heads.

Desist, oh brutal Queen; these statisticians are laudably trying to find a universally acceptable compromise. But psychologically a successful compromise must stand out as as a specially unique and convincing solution. Do any of these qualify?

Professor Bruno de Finetti (University of Rome): After a hasty reading of this beautiful paper, I suggest a modified version of Dickey's solution. After considering, for a broad range of choices, the pairs of priors and posteriors, choose that *pair* which gives the most subjectively satisfactory result. In other words, take an overall, rather than a one-sided, view of an acceptable choice.

Professor M. H. DeGroot (Carnegie–Mellon University, Pittsburgh): Congratulations to Professor Bernardo on an interesting and stimulating paper. Unfortunately, because of space limitations, I must skip over the many features that I liked and proceed directly to the few aspects with which I had difficulty. First, since the notion of reference posteriors depends on the idea of an infinite number of replications of ε, is this notion relevant to experiments that cannot be replicated?

Second, we cannot measure meaningfully the amount of information about θ in ε without considering the use to which this information is to be put. Choosing a measure of information is equivalent to considering a particular decision problem with a decision space D and loss function $L(\theta, d)$, as follows: For any density $p \in C$, let $U(p) = \min_d \int L(\theta, d)\, p(\theta)\, d\theta$ denote the uncertainty in p. Let p_k denote the posterior density $p(\theta \,|\, \mathbf{z})$, and let $E[U(p_k)] = \int U(p_k)\, p(\mathbf{z})\, d\mathbf{z}$. Then the expected information in $\varepsilon(k)$ is $I^\theta\{\varepsilon(k), p(\theta)\} = U(p) - E[U(p_k)]$, the expected reduction in uncertainty. Any such measure of information satisfies the properties of invariance, non-negativity, concavity and additivity mentioned by Professor Bernardo. He has taken U to be the entropy function H throughout his paper. Although H is useful in theory of communication, there is no compelling reason to restrict ourselves to this particular measure of uncertainty in statistical experiments. Using H is equivalent to considering a decision problem in which the statistician must choose a density function f from the class of all densities on Θ subject to the loss function

$$L(\theta, f) = -\log f(\theta).$$

The Bayes' decision is then to choose f to be the statistician's prior (or posterior) density p and the Bayes' risk $U(p)$ is just $H(p)$. The appropriateness of H in statistics is therefore no greater than the appropriateness of this decision problem.

The discussion in Section 3.4 suggests that it may be useful to partition prior distributions into three types: (i) proper priors, (ii) improper priors which must yield proper posteriors after some fixed number of observations and (iii) improper priors which do not satisfy (ii). In Section 3.4, Jeffreys' prior $\pi(\theta) \propto \{i(\theta)\}^{\frac{1}{2}}$ is proper. For the mean of a normal distribution, $i(\theta)$ is constant, and for the mean of a Poisson distribution, $i(\theta) = 1/\theta^{\frac{1}{2}}$. In each case, Jeffreys' prior belongs to category (ii). Haldane's prior in Section 3.4 belongs to category (iii).

Two final comments: (1) The rate at which the posterior distribution approaches normality seems to be irrelevant to the reference prior. Thus, at the end of Section 3.3 we could replace k in $\sigma^2(\tilde{\theta})/k$ by any function of k. Is this reasonable? (2) In Definition 1, the reference prior was obtained from the reference posterior. (Can we always obtain one?) But in Section 4, when nuisance parameters are present, the reference posterior is obtained from the reference prior. Is this switch necessary?

Dr A. W. F. Edwards (Gonville & Caius College, Cambridge): The first sentence of the paper contains the fallacy known to logicians as *petitio principii*, the fallacy of taking for granted a premiss which is equivalent to the conclusion. For although it might go without saying that the correct use of probability entails coherence, it does not go without saying that the correct medium for statistical inference is probability. This premiss is disputed.

Professor D. A. S. Fraser (University of Toronto): In this paper Professor Bernardo offers a thoughtful and comprehensive discussion within the Bayesian commitment. He acknowledges the familiar Bayesian difficulties involving reparameterization effects, marginalization paradoxes and strong inconsistency. He then confronts the prime Bayesian characteristic, that the results depend on the prior distribution. His approach is to seek a reference prior, "little relevant initial information" and to use the corresponding posterior directly or as a reference for other posteriors based on personal priors.

The marginalization paradoxes are avoided by a currently familiar procedure (Wilkinson, 1977), by making a virtue of a failure. The problem of inconsistent posteriors vanishes by having a wealth of priors and a corresponding compound wealth of posteriors. However, the procedure for component parameters does produce interesting and appealing results. It also raises the question as to what a distribution means if most of the probabilities cannot be used. In the extreme, each indicator parameter of a model, as a parameter of interest, could have its own prior and thus its own posterior probability: a prior for each possible posterior probability, conceivably all mutually inconsistent. The discrete example (coin) indicates the possibilities in this direction.

The author—within the Bayesian frame—focuses on the choice of a prior to describe "little relevant information". The difficulties lie in the commitment to the Bayesian frame; for some discussion see Fraser (1974).

Some recent research on information with and for statistical models (with D. Brenner, evolving from Fraser, 1972) leads to a classification of information as categorical, frequency and diffuse. Information can be available that is neither categorical nor frequency; the Bayesian approach makes no allowance for this, with resultant difficulties. Also, the proper classification for *no* information within a range is pure *categorical.* The Bayesian approach forces a measure on this range; the present paper attempts to minimize the effect.

To someone without a Bayesian commitment this thoughtful paper seems close to an interment of the Bayesian philosophy as an answer to statistics.

Professor S. Geisser (University of Minnesota): Inferential theories directed towards statements about parameters are largely irrelevant except they serve as a vehicle for theorists to beat one another over the head with. However, the notion of a reference prior is useful, not so much for the ostensible purpose intended—a statement about parameters—but as a device that permits the introduction of a predictive distribution of potential observables based principally on the observations at hand.

In a not too limited sense, the predictive distribution of a future observation is a surrogate for the sampling distribution of that observation, Geisser (1971). With this in mind, we outline another approach to the reference prior enigma.

Let $\mathbf{D}_N = (X_1, ..., X_N)$ represent a set of random variables which are to be observed and have joint density $f(\mathbf{d}_n \mid \theta)$, θ being an unknown set of parameters. Let a future random variable $\mathbf{X} \sim f(\mathbf{x} \mid \theta)$. For each $p(\theta)$ belonging to an admissible class C of prior densities for θ, a predictive density of $\mathbf{X}$ is obtained, say $g(\mathbf{x} \mid \mathbf{d}_N) = \int f(\mathbf{x} \mid \theta) p(\theta \mid \mathbf{d}_N)\, d\theta$ where $p(\theta \mid \mathbf{d}_N) = f(\mathbf{d}_N \mid \theta) p(\theta)/f(\mathbf{d}_N)$.

Using the Kullback and Leibler (1951) information distance (or some such other reasonable measure)

$$K(f, g \mid \theta, \mathbf{d}_N) = E_{\mathbf{x}\mid\theta}\left\{\log \frac{f(x \mid \theta)}{g(x \mid \mathbf{d}_N)}\right\}$$

and averaging over the sample space yields $M(f, g \mid \theta) = E_{\mathbf{D}_\mathbf{N}\mid\theta} K(f, g \mid \theta, \mathbf{D}_N)$. Then a reference prior (artifactual prior might be a better term, since the reference is to presumptive ignorance of an artifact of a statistical model imposed on the generation of data) could be defined as that member of the class C which minimizes $M(f, g \mid \theta)$ provided one exists, say $p^*(\theta)$ resulting in $g^*(\mathbf{x} \mid \mathbf{d}_N)$, for all admissible θ. The class C may actually be defined in a manner such that certain restrictions on the behaviour of $g(\mathbf{x} \mid \mathbf{d}_N)$ are incorporated. Such an approach has already been hinted at in Geisser (1971, 1977), Aitchison (1975) and Murray (1977).

Its first advantage is that predictive inference is stressed, not an irrelevant intermediary. Secondly, the question of nuisance parameters is avoided as interest is not focused on a marginal distribution of a particular set of parameters. Further, it also has a frequentist interpretation, if one prefers to think in those terms, in that g^* can be considered, for the given distance measure, as an optimal estimator of $f(\mathbf{x} \mid \theta)$ amongst estimators g generated by C.

All this is not meant to gainsay the interesting approach of Professor Bernardo which attempts to avoid some of the usual difficulties associated with the production of reference priors.

Professor I. J. Good (Virginia Polytechnic, U.S.A.): Professor Bernardo's paper is meaty but the central idea of considering the prior that "maximizes the missing information" was I think anticipated in Good (1969), to which Bernardo refers, and in Good (1968), where some of the results were announced. Those works mentioned among other things that (a) the concept of the "utility" or "quasi-utility" of a distribution merited more attention (see also Good, 1960); (b) when the distribution is parameterized this amounts to talking about the utility of assigning values to the parameters; (c) when one has such measures of utility, the minimax prior for the parameters is known, by Wald's theorem, to be one of smallest prior utility; (d) although minimax priors have disadvantages they have nice invariant properties; (e) an interesting quasi-utility is an information measure or expected weight of evidence as in Bernardo's work and as used by Turing (see Good, 1979 for more history); (f) in this case the minimax prior is the least informative one and could be called the *minimax-information* or *minimax-evidence* prior; (g) this is the Jeffreys' prior in the continuous case; (h) "the 'least favourable' initial distribution, if it exists ... is ... invariant .... It generalizes (i) the Jeffreys–Perks invariance theory; (ii) a principle of minimum discriminability for determining a distribution (Kullback, 1959); and (iii) the similar principle for maximum entropy for initial distributions (Jaynes, 1957)."

I was pleased to see the concept of minimax-information priors developed so well in so many directions by Bernardo. In particular I was intrigued by his proposal in Section 6 for applying the idea to hyperpriors, for I have been advocating hierarchical Bayesian methods for a very long time (Good, 1952), especially in connection with multinomials and contingency tables (for example, Good, 1965, 1967, 1976; Good and Crook, 1974). In these applications improper hyperpriors, such as the Jeffreys–Haldane prior, cannot be used but can be approximated so as to model "ignorance". It might be interesting to consider the minimax-information hyperpriors for these applications and to compare the effects with those of the log-Cauchy hyperpriors that I used.

Professor J. A. Hartigan (Yale University): Although Professor Bernardo states that "much attention to mathematical detail would be premature", my own belief is that the difficulty with "improper" priors is one of mathematical detail. For this reason, Definition 1 of a reference prior, in which a limiting notion is used to avoid explicit handling of improper priors, demands careful attention to mathematical detail. What is the role of the compact class C of admissible priors which is never mentioned in later derivations? It does not appear that compactness of C in the topology of weak convergence of priors is sufficient to "guarantee the existence of the maximum". Consider Example 3.5; if the prior density is p, and prior distribution function is P,

$$I^{\theta}(\varepsilon, P) = -\int \{P(x+\tfrac{1}{2})-P(x-\tfrac{1}{2})\} \log\{P(x+\tfrac{1}{2})-P(x-\tfrac{1}{2})\}\, dx$$

which is infinite for the discrete prior $P(\theta = k) = \alpha/k(\log k)^2$. In general $I^{\theta}(\varepsilon(k), P)$ will be infinite, for all k, for the prior density

$$p(\theta) = \lambda/\theta(\log \theta)^2 \quad \theta>2,$$
$$p(\theta) = 0, \qquad\qquad \theta<2.$$

It will be infinite for many proper priors with the appropriate tail behaviours. The same sort of degeneracy applies to normal location. Expected increase of information is technically inadequate as a criterion for evaluating priors, since so many give infinite increases in information.

Professor S. James Press (University of California): The author has written what I feel is an important paper in the field of Bayesian inference. I am sure it will be cited on numerous occasions in the future, and will be used by many research workers in this field to justify the prior distributions that they use.

The principal concern I have with the paper revolves around the nature of the approximations used by the author to develop sequences of prior distributions approaching the reference prior. The quality and implications of the approximations are not clear. What constraints are being imposed on the analyst's subjective beliefs by this approximation?

The notion of using Jeffreys' invariant prior for simultaneous inference about all parameters is useful. It is also useful for the author to point out the importance of distinguishing between the quantity of interest and the complete parameter (in the final paragraph before Section 4.1). It should be noted that Professor Arnold Zellner and I have made this distinction also, in our paper on the posterior distribution of the multiple correlation coefficient (1978). There, we showed that if this distinction is ignored, the likelihood function will not be the correct one to use and an apparent paradox arises. The way the problem is resolved is consistent with the argument of Jaynes (1978), which is supported by the approach of the present paper.

I congratulate the author on making inroads on a difficult, but very important problem.

Dr A. M. Skene (University of Nottingham): I find this paper interesting and I applaud its objective. However, I would be grateful for some further explanation on one matter and I wish to offer an additional comment.

Consider the balanced one-way random effects model, (see, for example, Box and Tiao, 1973, p. 244). An experimenter approaches me with suitable data and expresses an interest in the overall mean and both variance components. I duly supply him with the *joint* posterior reference distribution. Can the experimenter use this to calculate the *marginal* distribution for, say, the within groups variance? If so, what, if anything, is he allowed to infer from that margin? I feel that there exists a practical distinction between those parameters which are deemed to be nuisance parameters, *a priori*, and those which are temporarily so designated while investigating individual margins of a joint posterior density.

The concept of a reference posterior distribution forces one to consider the consequences of different choices of prior. In a somewhat cursory numerical investigation which set out to show that the choice of prior was immaterial given sufficient data, marginal distributions for the two variance components of the model previously mentioned were plotted for six different choices of prior. Let σ^2 and σ_a^2 be the within and between groups variance components having unbiased estimates S^2 and S_a^2 respectively. The priors chosen were (i) uniform, (ii) $\sigma^{-2}(\sigma^2+J\sigma_a^2)^{-1}$, where J is the number of observations per group, (iii) $\sigma^{-2}. \sigma_a^{-2}$, (iv) independent χ_1^{-2} densities whose modes coincided with the unbiased estimates, (v) independent χ_1^{-2} densities whose modes differed from the unbiased estimates, (vi) $\sigma^{-2} \exp(-S^2/\sigma^2). \sigma_a^{-2} \exp(-S_a^2/\sigma_a^2)$, the last being the product of two improper limiting distributions obtained from the χ^{-2} distribution. The two examples illustrated in Figs D2 and D3 are due to Box and Tiao (1973, p. 246) and Hill (1976), respectively. In view of

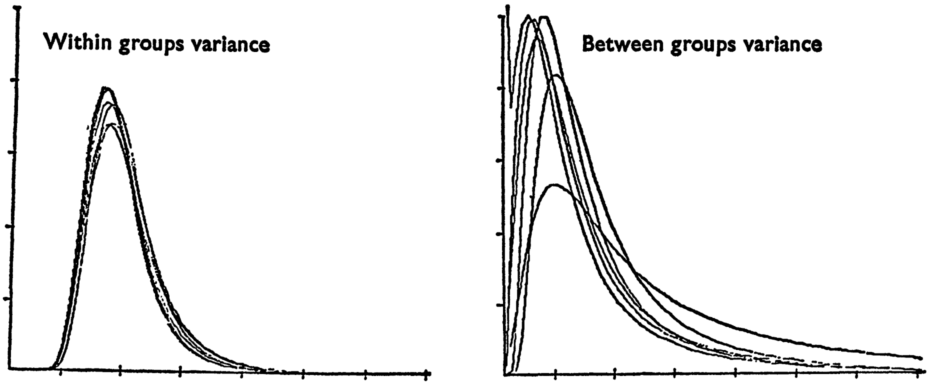

Fig. D2. Variance components of a one-way Anova having six groups and five observations per group

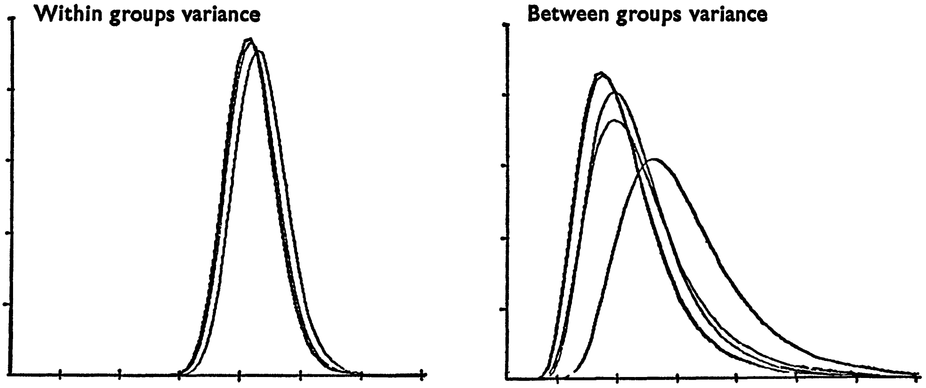

Fig. D3. Variance components of a one-way Anova having twenty groups and ten observations per group

the marked effects of these priors, all of which purport to provide little information, a major question which is still to be answered is simply "when, for the purposes of inference, is one curve a valid approximation for another?"

The Author replied later, in writing, as follows.

I am most grateful to the contributors for interesting and very useful comments. I have to thank most of them for their generally warm and encouraging tone, despite of this Societys' reputation as a forum for violent exchanges; for providing further insight into the consequences of the procedure proposed and for suggesting new problems for research. In the following, I shall try to give an answer to the queries which have been raised.

Professor Copas is certainly right when he mentions that the expected information I^θ defined by (1) depends only on the shape of the distribution and is independent of the actual values of the variables. He goes on to describe a situation in which the *risk* involved in *decision*-making is higher when the probability distribution is the same. However, he misses that, in this paper, we are facing a *purely inferential* case, where one is only interested in gaining knowledge about the parameter of interest and has no specific decision in mind. The connection between the logarithmic measure of information and scientific inference has been established elsewhere (Bernardo, 1979) within a Bayesian framework. The reason why the method works, as illustrated by the examples of Sections 3.1, 3.5 and 4.1 and by that provided by Dr O'Hagan in the discussion, is *not* asymptotic normality but the deep connection between scientific inference and Shannon's information measure.

It must be stressed, however, that the relevant quantity is information *not* entropy. For, the axiomatic justification of entropy does not extend to the continuous case and, moreover, for a continuous random quantity X, $H\{p(x)\} = -\int p(x) \log p(x)\, dx$ is *not* the limit of discrete entropies, is *not* invariant under one-to-one transformations of X and has *no* precise meaning as a measure of the uncertainty attached to X. However, $I^x\{Y, p(x)\} = H\{p(x)\} - E_y\, H\{p(x \mid y)\}$ *is* invariant under one-to-one transformations of X and *is* a measure of the amount of information that Y is expected to provide about the value of X with a precise interpretation in terms of the expected number of questions about X that it would be necessary to ask to obtain the same level of knowledge as that expected to be provided by Y (Rényi, 1970, p. 564). I believe I^θ relates very directly to statisticians' ideas of information and inference although certainly not to those of risk and decision.

It seems rather likely that if one had a prior knowledge of the form $p(\theta \in S) = p$, one would also have a similar knowledge of the form $p(\theta \in S_i) = p_i$ for other sets S_i close to S. It is clear that a number of statements of this form will produce a more smooth reference prior. In any case, the object of discussion and interpretation should be the reference posterior, not the operational prior.

A situation in which $p(\theta \in S) = p$ may be the only prior knowledge available is when $\{\theta \in S\}$ is a consequence of some scientific theory, the prior probability of which is p. This really occurs when the parameter of interest is not θ but $\psi(\theta)$ defined by $\psi(\theta) = 1$ if $\theta \in S$, $\psi(\theta) = 0$ otherwise, with $p(\psi = 1) = p$ and $p(\psi = 0) = 1-p$. The operational prior is then that quoted by Professor Copas and the resulting reference posterior for the parameter of interest is such that

$$\frac{p(\psi = 1 \mid x)}{p(\psi = 0 \mid x)} = p \int_S p(x \mid \theta)\, i(\theta)^{\frac{1}{2}}\, d\theta \Big/ (1-p) \int_{\bar{S}} p(x \mid \theta)\, i(\theta)^{\frac{1}{2}}\, d\theta$$

which is, I feel, a rather sensible result.

Reference posteriors are not inferences in their own right in the sense that they do not describe the scientist's personal opinions but those of someone with a very special kind of knowledge: that which leaves most to be said by the data. However, reference posteriors may certainly be updated in a sequential way as new information arises from the same model. Indeed, the operational prior depends on the model only through the asymptotic posterior of its parameters which is, of course, independent of the particular sample one might have obtained.

Finally, I hope that reference posteriors will not prove to be so dangerous as Professor Copas fears. The proposal is very simple: people should quote both the personal and the reference posterior and explain that the discrepancy among them is solely due to the prior knowledge they had about the parameters.

Dr O'Hagan insists on treating operational priors as a representation of ignorance and proceeds to measure this ignorance by using entropies. I have just mentioned that operational priors are *not* representations of ignorance but approximate descriptions of a very specific type of knowledge: that in which most remains to be learned from the data about the parameter of interest θ. No wonder that the description of such knowledge depends on the choice of the parameter of interest; the entropy of the resulting prior is thus irrelevant. He goes on to mention that the Bayesian argument acknowledges the fact that a single toss of the coin tells us nothing about whether it is fair, by the prior and posterior probabilities that the coin is fair being equal; but the fact remains that $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ is the *only* prior which produces the reference posterior $(\frac{1}{2}, \frac{1}{2})$ one would expect to obtain.

I have tried to stress that a reference posterior is an *origin* for *inference*. It should be used to measure the relative importance of prior knowledge, but it would certainly be foolish to use it to take a personal decision in *lieu* of the personal posterior which describes the decision-maker's opinions. Only if, by some personal or political reason, one wanted to justify a decision in terms of some agreed initial knowledge and some data, one could use the reference posterior compatible with such knowledge. In this case, one should identify the utility function, treat as parameter of interest the parameters from which this function depends and use a reference posterior for them to obtain a *reference decision*. The procedure will be coherent in that one would be adopting a reference posterior as a personal one and acting accordingly. In Dr O'Hagan's example, the decision problem (to accept or not a set of bets) involves all the parameters. I would decide on the basis of my personal opinions about them, which only in very special circumstances would be described by the corresponding reference posterior $(\frac{1}{3}, \frac{2}{3}, 0)$. Clearly, to use this distribution without good reason may well be foolish, but it is not incoherent; I am afraid that Dr O'Hagan will not become rich by being my bookmaker. I will outline later a procedure to produce reference posteriors that could be used in decision-making to produce reference decisions; this takes into account the utility function of the decision problem.

In his last point, Dr O'Hagan is interested in the probability of the next toss of the counterfeit coin being heads. Thus, the quantity of interest is whether the next toss is heads ($y = 1$) or not ($y = 0$). It will be shown later that the operational prior required to obtain reference predictive distributions is precisely that required to obtain the joint reference posterior for all the parameters in the model. In this example, such operational prior is obviously $\pi_\psi(\psi) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$; the corresponding reference posterior distribution after one toss of the coin resulting in heads is $\pi(\psi \mid x = 1) = (\frac{1}{3}, \frac{2}{3}, 0)$ and thus the reference predictive probability desired $\pi(y = 1 \mid x = 1) = \frac{5}{6}$ and certainly *not* 1·05!

In conclusion, although some obscurities may remain in the construction of what Dr O'Hagan calls conditional reference priors, I do not think his testimony provides evidence against them. Moreover, to judge them fairly, one should keep in mind the rather appealing solutions obtained with their use to the problems discussed in Section 5.

Professor A. F. M. Smith complains that further light is not shed on the problems raised by Stein (1956) and Stone (1976). With respect to Stein's inadmissibility result I have two comments: (i) Reference posteriors are intended to be an origin for admissible inferences and must therefore produce results which are arbitrarily close to admissibility but are not necessarily admissible; although the mean $\bar{x}$ of the reference posterior for the means of a multivariate model is an inadmissible estimator for dimensions larger than two, it is arbitrarily close to estimators of the form $\alpha\bar{x} + (1-\alpha)\mu_0$, which are posterior means for suitably chosen proper priors and are, therefore, admissible. (ii) Indeed, one may have a prior knowledge about some kind of relation among the μ_i's; in this case, one may obtain the reference posterior in the restricted class of priors compatible with such assumption and this will produce admissible estimators of the type mentioned above.

I have not included a discussion of Stone's examples of strong inconsistency to keep the paper within reasonable size, but it may be verified that maximization of the missing information produces sensible answers in the two examples discussed in that paper. Indeed, in the Flatland example there is a relevant nuisance parameter ω, namely the position of the woman and the soldier just before leaving the treasure, which was ignored in Stone's "Bayesian" analysis. Thus, the data x will consist on the direction in which the thread is pointing from the endpoint (N, S, E, W). The parameter of interest θ concerns the position of the treasure relative to the endpoint (N, S, E, W) and the nuisance parameter ω the position of the woman and the soldier one step before leaving the treasure, relative to θ, $\{\mathrm{N}(\theta), \mathrm{S}(\theta), \mathrm{E}(\theta), \mathrm{W}(\theta)\}$. Clearly, the likelihood of, say $x = \mathrm{E}$ is given by

$$\begin{aligned}
&p\{x = \mathrm{E} \mid \theta = \mathrm{E},\ \omega = \mathrm{W}(\theta)\} = 0, && p\{x = \mathrm{E} \mid \theta = \mathrm{S},\ \omega = \mathrm{N}(\theta)\} = \tfrac{1}{3},\\
&p\{x = \mathrm{E} \mid \theta = \mathrm{E},\ \omega \neq \mathrm{W}(\theta)\} = 1, && p\{x = \mathrm{E} \mid \theta = \mathrm{S},\ \omega \neq \mathrm{N}(\theta)\} = 0,\\
&p\{x = \mathrm{E} \mid \theta = \mathrm{N}, \omega = \mathrm{S}(\theta)\} = \tfrac{1}{3}, && p\{x = \mathrm{E} \mid \theta = \mathrm{W}, \omega = \mathrm{E}(\theta)\} = \tfrac{1}{3},\\
&p\{x = \mathrm{E} \mid \theta = \mathrm{N}, \omega \neq \mathrm{S}(\theta)\} = 0, && p\{x = \mathrm{E} \mid \theta = \mathrm{W}, \omega \neq \mathrm{E}(\theta)\} = 0.
\end{aligned}$$

Using the results in Section 3.1, the missing information about θ is maximized when

$$\pi\{\theta = \mathrm{E}\} = \pi\{\theta = \mathrm{N}\} = \pi\{\theta = \mathrm{W}\} = \pi(\theta = \mathrm{S}\} = \tfrac{1}{4}$$

and the missing information about ω given θ when

$$\pi\{\mathrm{E}(\theta) \mid \theta\} = \pi\{\mathrm{N}(\theta) \mid \theta\} = \pi\{\mathrm{W}(\theta) \mid \theta\} = \pi\{\mathrm{S}(\theta) \mid \theta\} = \tfrac{1}{4}.$$

Using this operational prior, the reference posterior for the parameter of interest is clearly

$$\pi\{\theta = \mathrm{E} \mid x = \mathrm{E}\} = \tfrac{3}{4}, \quad \pi\{\theta = \mathrm{S} \mid x = \mathrm{E}\} = \pi\{\theta = \mathrm{W} \mid x = \mathrm{E}\} = \pi\{\theta = \mathrm{N} \mid x = \mathrm{E}\} = \tfrac{1}{12}$$

in agreement with the coverage probabilities. Moreover, as Professor Stone mentions in his reply to the discussion, the reason for the strong inconsistency in his Example *B* is surely the non-identifiability of θ. But, as Dr O'Hagan has just remaked, the procedure described in this paper makes explicit the inexistence of a reference posterior for a non-identifiable parameter by producing an arbitrary operational prior.

Professor Smith is right when he notes that in the presence of several nuisance parameters, say $\omega = (\omega_1, \omega_2)$, the joint reference conditional prior $\pi(\omega_1, \omega_2 \mid \theta)$ might be different from the product $\pi(\omega_1 \mid \omega_2, \theta)\, \pi(\omega_2 \mid \theta)$; this was to be expected since, given θ, the first alternative is equivalent to a situation in which the parameter of interest is (ω_1, ω_2) while the second is equivalent to one in which ω_1 is the parameter of interest and ω_2 a nuisance parameter. The choice among them will depend on the type of reference knowledge one wants to describe.

Finally, Professor Smith mentions the truly important class of problems which arise in Bayesian choice of model, when one is interested in the posterior probabilities of a list of alternative models with possibly different dimensionalities. I have reasons to believe that the method of maximizing the missing information does indeed produce sensible answers in this area too. However, the topic is much too vast to be covered here; I hope to be able shortly to report on it elsewhere.

Professor Dawid asks whether the procedure described may be extended to produce *reference decisions*. I think this can be done. Consider a decision problem (D, Θ, u) where D is the decision space, Θ the parameter space and $u(d, \theta)$ the utility function. Let x be the result of some experiment ε and $\mathbf{z} = \{x_1, ..., x_k\}$ that of k independent replications of ε. The expression

$$\int \max_d \int u(d, \theta) p(\theta \mid x)\, d\theta p(x)\, dx - \max_d \int u(d, \theta) p(\theta)\, d\theta \tag{1}$$

then measures the increase in utility to be expected from performing ε. A measure of the *missing utility* that could eventually be provided by infinite replications of ε is

$$\lim_{k\to\infty} \int \max_d \int u(d, \theta) p(\theta \mid \mathbf{z})\, d\theta p(\mathbf{z})\, dz - \max_d \int u(d, \theta) p(\theta)\, d\theta \tag{2}$$

which, under suitable conditions, will be simply

$$\int \{\max_d u(d, \theta)\} p(\theta)\, d\theta - \max_d \int u(d, \theta) p(\theta)\, d\theta, \tag{3}$$

i.e. the expected value of perfect information. The prior $\pi(\theta)$ which maximizes (2) within the class C of admissible priors may be seen as a reference prior for the decision problem considered, in that it leaves most to be *gained* from the data. The optimal decision attached to the corresponding reference posterior would be a suitable *reference decision* to be compared with the optimal decisions attached to personal posteriors.

A very simple example is provided by the decision problem of estimation with quadratic loss. Here, $D = \Theta$ and $u(d, \theta) = -A(d-\theta)^2$; the first integral in (2) vanishes and the second term is the prior variance of θ. Thus, the *reference estimator* with quadratic loss is the posterior mean corresponding to that prior with larger variance among those compatible with the assumptions made. It is apparent from this example, that reference decisions are not necessarily unique although, often, sensible restrictions in the class of admissible priors will imply uniqueness.

It may be verified that the method proposed in this paper to derive reference posteriors is the special case of the procedure outlined above, where D is the class of distributions of θ (so that we are in a problem of *pure inference*) and the utility function is of the form

$$u(d, \theta) = A \log p(\theta) + B(\theta). \tag{4}$$

Indeed, in this case expression (1) becomes $I^\theta\{\varepsilon, p(\theta)\}$, and maximizing the missing utility means maximizing the missing information. The rationale for using the particular utility function (4) may be found in Bernardo (1979).

In the second part of his remarks, Professor Dawid worries about the procedure to obtain a reference posterior probability for the event that the parameter of interest θ belongs to a given set S. As he mentions, one could define a new parameter of interest ξ such that $\xi = 1$ if $\theta \in S$ and $\xi = 0$ otherwise and determine its reference posterior distribution or, alternatively, to compute the probability $\int_S \pi(\theta \mid x)\, d\theta$ attached to S by the reference posterior distribution. It may be seen however that both methods give the same result.

Indeed, the reference conditional prior $\pi(\theta \mid \xi)$ is, as mentioned by Professor Copas, of the form

$$\pi(\theta \mid \xi = 1) = c_1 \pi(\theta), \quad \text{if } \theta \in S, \qquad \pi(\theta \mid \xi = 0) = c_2 \pi(\theta), \quad \text{if } \theta \notin S,$$

where $\pi(\theta)$ is the reference prior for θ.

$$c_1^{-1} = \int_S \pi(\theta)\, d\theta \text{ and } c_2^{-1} = \int_{\bar{S}} \pi \ (\theta)\, d\theta.$$

Moreover, the reference (unconditional) prior for ξ is found to be $\pi(\xi = 1 | = c_1^{-1}$ and $\pi(\xi = 0) = c_2^{-1}$. It follow that

$$\pi(\xi = 1 \mid x) = \int_S p(x \mid \theta)\, \pi(\theta)\, d\theta; \;\; \pi(\xi = 0\, (x \mid = \int_{\bar{S}} p(x \mid \theta)\, \pi(\theta)\, d\theta$$

and therefore $\pi(\xi = 1 \mid x) = \int_S \pi(\theta \mid x)\, d\theta$ as desired. Consequently, one may integrate in reference posterior distributions to obtain reference posterior probabilities.

Professor Bartholomew wonders whether the procedure presented may be applied with sequential sampling schemes where the stopping rule depends on the observations obtained to date. I do not see why not. The likelihood of the result finally obtained when the experiment comes to an end will be of the form

$$p(\mathbf{x} \mid \theta, \omega, \tau) = f(\mathbf{x} \mid \theta, \omega)\, g(n \mid \mathbf{x}, \theta, \omega, \tau), \tag{5}$$

where $\mathbf{x} = \{x_1, \ldots, x_n\}$, $f(\mathbf{x} \mid \theta, \omega)$ is the probability of obtaining the sample $\mathbf{x}$ given θ, ω and n and $g(n \mid \mathbf{x}, \theta, \omega, \tau)$ the probability of stopping there after observing $\mathbf{x}$. Thus, the only consequence of the stopping rule is the introduction of the new nuisance parameter τ, and the methodology described may be used to obtain the reference posterior for the parameter of interest θ.

On his second point, Professor Bartholomew is certainly right when he mentions that the parameter of interest is often a future observation; a simple example of this was mentioned by Dr O'Hagan in the discussion. According to the procedure proposed in the paper, if the parameter of interest is a future observation y from $p(y \mid \theta)$, the operational prior for the parameter of interest, i.e. the predictive distribution $\pi(y) = \int p(y \mid \theta)\, \pi(\theta)\, d\theta$ should be one maximizing the missing information about y and, among those prior distributions of θ which satisfy this condition, one should select that maximizing the missing information about the nuisance parameter, i.e. the missing information about θ. We shall now show that the result of such a programme is simply the operational prior for θ, $\pi_\theta(\theta)$.

Indeed, if $\mathbf{z} = \{x_1, \ldots, x_k\}$, we have by definition

$$I^y\{\varepsilon(k), p(\theta)\} = \int p(\mathbf{z}) \int p(y \mid \mathbf{z}) \log \frac{p(y \mid \mathbf{z})}{p(y)}\, dy\, d\mathbf{z}.$$

Under regularity conditions, for large k we have $p(y \mid \mathbf{z}) = p(y \mid \hat{\theta})$ with $\hat{\theta}$ in a neighbourhood of θ. Thus,

$$\begin{aligned} I^y\{\varepsilon(k), p(\theta)\} &= \int p(\mathbf{z}) \int p(y \mid \hat{\theta}) \log \frac{p(y \mid \hat{\theta})}{p(y)}\, dy\, d\mathbf{z} + O(1) \\ &= \int p(\theta) \int p(\mathbf{z} \mid \theta) \int p(y \mid \hat{\theta}) \log \frac{p(y \mid \hat{\theta})}{p(y)}\, dy\, d\mathbf{z} + O(1) \\ &= \int p(\theta) \int p(y \mid \theta) \log \frac{p(y \mid \theta)}{p(y)}\, dy\, d\theta + o(1) \quad (k \to \infty). \end{aligned} \tag{6}$$

But, if we write $p(y) + \alpha\delta(y)$ in place of $p(y)$, a necessary condition for $p(y)$ to be an extreme of (6) such that $\int p(y)\, dy = 1$ is, using Lagrange multipliers, that

$$\int p(\theta) \int \left\{ \frac{p(y \mid \theta)}{p(y)} - \lambda \right\} \delta(y)\, dy\, d\theta = 0 \quad \text{for any } \delta(y).$$

This implies $\int p(\theta)\, p(y \mid \theta)\, d\theta = \lambda p(y)$, which is simply the definition of the predictive density $p(y)$. Thus, the first condition on $\pi(\theta)$ turns out to be vacuous, and one must simply maximize the missing information about θ to obtain the operational prior for θ, $\pi_\theta(\theta)$, which will therefore be the prior required to produce reference predictive distributions.

In the case of Bernoulli observations, the operational prior is $\mathrm{Be}\,(\theta \mid \frac{1}{2}, \frac{1}{2})$ and, thus, the reference predictive probability of obtaining a new success if one has previously obtained r successes out of n trials is $\pi(y = 1 \mid r, n) = (r + \frac{1}{2})/(n+1)$.

Professor Lindley is worried by the dependence of the reference posterior on the sampling rule in violation of the likelihood principle; I must admit that I was puzzled myself when I first realized this. However, when one looks more closely into the problem of scientific reporting, one realizes that scientists are usually required by their colleagues to specify not only their results but also the conditions in which the experiment has been performed, i.e. the *design* of the experiment; if I am right, they *should* be asked to do so. Indeed, it is known that, even from a purely personalistic point of view, one *must* integrate over the sample space to design an experiment. It does not seem unnatural to me that one has to do the same to analyse the implications of its results.

In Haldane sampling, when one is sampling until r successes are obtained, one is somehow assuming that r successes *will* eventually appear, a different situation from that in ordinary sampling. This is duly reflected in the reference posterior for Haldane sampling,

$$\pi(\theta \mid n) \propto \{\theta^2(1-\theta)\}^{-\frac{1}{2}} p(r \mid \theta) \propto \mathrm{Be}\,(\theta \mid r, n-r+\tfrac{1}{2})$$

which is *not* proper if $r = 0$. This is only natural, for we are assuming that a success *will* appear and, in the absence of other information, we cannot make inferences otherwise. However, with ordinary sampling, we have the situation described in Section 3.4 and we can make inferences even if $r = 0$. I find these results quite reasonable, and I would suggest that, indeed, scientific reporting on the implications of some experimental results requires the knowledge of their design.

In his second example, Professor Lindley proposes to obtain the reference prior for (θ, λ) after $x = (D, W, S)$ have been observed when

$$p(D, W, S \mid \theta, \lambda) = \frac{\Gamma(D+W+S)}{\Gamma(D)\,\Gamma(W)\,\Gamma(S)} \{\lambda[1-(1-\delta)\,\theta]\}^D\{(1-\delta\lambda)\,\theta\}^W\{(1-\lambda)\,(1-\theta)\}^S$$

and δ is a known constant, i.e. a trinomial with $p_1 = (1-\delta\lambda)\,\theta$, $p_2 = (1-\lambda)\,(1-\theta)$ and $p_3 = 1-p_1-p_2$. A straightforward extension of the results in Section 3.3 shows that, under regularity conditions, the reference prior to make inferences about a vector θ with no nuisance parameters is Jeffrey's multivariate $|\mathbf{F}(\theta)|^{\frac{1}{2}}$ where $\mathbf{F}(\theta)$ is Fisher's information matrix. In the trinomial case with parameters p_1 and p_2, it is easily verified that

$$\mathbf{F}^{-1}(p_1, p_2) = \begin{bmatrix} p_1(1-p_1) & p_1 p_2 \\ p_1 p_2 & p_2(1-p_2) \end{bmatrix}. \tag{7}$$

Moreover, if $\xi = \xi(\theta)$ is a one-to-one transformation of θ, it is easily established that the corresponding information matrix $\mathbf{F}(\xi)$ is related to $\mathbf{F}(\theta)$ by the equation

$$\mathbf{F}^{-1}(\xi) = (\nabla\xi)^{-1}\,\mathbf{F}^{-1}(\theta)\,\{(\nabla\xi)^{\mathrm{T}}\}^{-1}, \tag{8}$$

where $(\nabla\xi)$ is the square matrix of typical element $\partial\xi_i/\partial\theta_j$. Thus, using (7) and (8), the inverse of the information matrix of (θ, λ) is

$$\mathbf{F}^{-1}(\theta, \lambda) = \mathbf{T}\mathbf{F}^{-1}(p_1, p_2)\mathbf{T}^{\mathrm{T}},$$

where

$$\mathbf{T}^{-1} = \begin{bmatrix} 1-\delta\lambda & -\delta\theta \\ -(1-\lambda) & -(1-\theta) \end{bmatrix},$$

$\mathbf{F}^{-1}(p_1, p_2)$ is given above, $p_1 = (1-\delta\lambda)\,\theta$ and $p_2 = (1-\lambda)\,(1-\theta)$. The reference prior for (θ, λ) is then

$$\pi(\theta, \lambda) \propto |\mathbf{F}^{-1}(\theta, \lambda)|^{-\frac{1}{2}}.$$

After some rather tedious algebra, this becomes

$$\pi(\theta, \lambda) = \{\lambda(1-\delta\lambda)\,(1-\lambda)\}^{-\frac{1}{2}}[\{1-(1-\delta)\,\theta\}\,\theta(1-\theta)]^{-\frac{1}{2}}$$

which *does* factorize.

Moreover, the reference prior for λ, i.e.

$$\pi(\lambda) \propto \{\lambda(1-\delta\lambda)\,(1-\lambda)\}^{-\frac{1}{2}}$$

does reduce to $\lambda^{-\frac{1}{2}}(1-\lambda)^{-\frac{1}{2}}$ when $\delta = 0$. If $\delta = 1$, it reduces to $\lambda^{-\frac{1}{2}}(1-\lambda)^{-1}$, i.e. the reference prior for Haldane sampling with $\theta = (1-\lambda)$; I suspect that $\delta = 1$ is a limiting condition which precisely implies this type of sampling rule.

Dr Brown proposes to consider a problem of diagnosis where the new undiagnosed case is known to belong to one of two multinomial populations. As he mentions, the problem is a specific example of the model choice problem mentioned by Professor A. F. M. Smith. In Dr Brown's problem, the likelihood of the symptoms x observed in the new case is of the form

$$p(x \mid \theta, \phi, \delta = 1) = \prod_{i=1}^{n} \theta_i^{x_i}, \quad p(x \mid \theta, \phi, \delta = 0) = \prod_{i=1}^{n} \phi_i^{x_i}$$

and one is interested in the reference posterior probability of δ after some training data z and the symptoms x of the new case have been observed.

Since the parameter of interest is not (θ, ϕ) but δ, while $\omega = (\theta, \phi)$ is a set of nuisance parameters, the appropriate operational prior is not a Dirichlet but one of the form $\pi(\delta)\,\pi(\omega \mid \delta)$ where $\pi(\delta)$ maximizes the missing information about δ and $\pi(\omega \mid \delta)$ the missing information about ω given δ. I have not worked out the details, but I would expect $\pi(\delta)$ to depend on n in a way that will avoid the problems pointed out by Lindley.

Professor C. A. B. Smith is to be congratulated for his discovery of a missing page of *Alice in Statland* which we have all enjoyed so much. I am afraid, however, that the author of the book did not transcribe properly the conversation between Alice and the Better Bernardians' Enquiry Office. Indeed, in the internal report from that conversation, I have found that what Alice was told was that the mean effect of the magnetic field comes to 16 grams whether you are interested or not in the variability as well, and that the concept of information may well be taken as primitive and probability derived from it.

Professor de Finetti suggests that one could assess subjectively whether the posterior distribution gives a satisfactory result. I believe one should certainly do that: how would one deal otherwise with, say, totally unexpected results? I only hope that some people will find subjectively satisfactory reference posterior distributions as a description of the knowledge provided by the data.

Professor De Groot wonders whether the notion of reference posteriors is relevant to experiments that cannot be replicated. Since only a formal, *conceptual* replication of the experiment *performed* is necessary, and one can always imagine this, I think the procedure may be used to analyse the result of any experiment.

I certainly agree with Professor De Groot's general definition of information; this, in turn, is a special case of the approach to reference decisions I have outlined above. Indeed, the appropriateness of Shannon's information measure in statistics is not greater than that of the utility function $u(p, \theta) = \log p(\theta)$; but, possibly, this *is* the utility function appropriate to scientific inference (Bernardo, 1979).

The idea of a reference posterior is based on some sort of "measure" of the "distance" between between prior and perfect knowledge; I do not see why this distance should depend on the rate at which perfect knowledge may be obtained any more than the distance between Pittsburgh and Valencia should depend on the speed at which my friend could come to visit me.

A formal definition of the operational prior when nuisance parameters are present should also be given in terms of the reference posterior to avoid convergence problems. Thus, the operational prior would be that function $\pi_\theta(\theta, \omega)$ which produces, via Bayes' theorem, the posterior

$$\pi(\theta \mid x) = \lim p_k(\theta \mid x),$$

where

$$p_k(\theta \mid x) \propto \pi_k(\theta) \int p(x \mid \theta, \omega)\, \pi_k(\omega \mid \theta), d\omega,$$

where $\pi_k(\omega \mid \theta$ maximizes $I^{\omega/\theta}\{\varepsilon(k), p(\omega \mid \theta)\}$ and $\pi_k(\theta)$, maximizes $I^\theta\{\varepsilon(k), p(\theta)\, \pi_k(\omega \mid \theta)\}$, within the class of admissible priors. Under mild regularity conditions, the reference posterior will always exist, since the maxima exist by the concavity of the information measure as a functional of the prior and their limit by the asymptotic convergence of posterior distributions.

Dr Edwards does not comment on the paper: he simply refuses to accept the Bayesian approach to inference. I do not think this is the best occasion to discuss foundations, but I would like to see Dr Edwards' explicit solutions to any of the problems mentioned in Section 5, and I would like to know whether he claims them to be better in any well-defined sense.

Professor Fraser wonders what a distribution means if, according to him, most of the probabilities cannot be used. As I have mentioned before, in reply to Professor Dawid, *all* the *relevant* probabilities, i.e. the posterior probabilities of the parameter of interest belonging to any set, may be used and are consistent with those obtained when their respective indicators are considered as parameters of interest.

I am afraid that I cannot agree with Professor Fraser's apocalyptic conclusion. Indeed, the standard arguments against Bayesian methods focus on their dependence on prior opinions which somehow conflict with scientific reporting, and reference posteriors provide a procedure to bypass this problem.

Professor Geisser stresses quite properly the practical importance of prediction and goes on to propose as an operational prior, when interest is in prediction of the next observation x, that which, for every θ, minimizes

$$\int p(\mathbf{z} \mid \theta) \int p(x \mid \theta) \log \frac{p(x \mid \theta)}{p(x \mid \mathbf{z})} dx\, d\mathbf{z} \tag{9}$$

where $\mathbf{z} = \{x_1, ..., x_n\}$. This is the expected value of the amount of information about x which perfect knowledge about θ would provide over and above that contained in $p(x \mid \mathbf{z})$; it is therefore a non-negative quantity whose minimum value, zero, is attained for each θ, when $p(x \mid \mathbf{z}) = p(x \mid \theta)$.

I do not think this procedure will produce sensible answers, if only because the prior which minimizes (9) will generally depend on θ. Moreover, the results obtained may be far from satisfactory. To see this, consider the problem introduced by Dr O'Hagan in the discussion where one is interested in predicting the second toss from the counterfeit coin of Section 4.1. Expression (9) is then minimized by $\pi(\psi) = (1, 0, 0)$ if $\psi = \psi_0$ (fair coin) and by any prior of the form $\pi(\psi) = (0, p, 1-p)$ if ψ is either ψ_1 or ψ_2. I do not know how Professor Geisser would choose among those priors, but I suspect that he would not like to use either of them.

I certainly agree with Professor Geisser in that when one is interested in prediction the question of nuisance parameters is irrelevant. But, as I have shown above in reply to Professor Bartholomew, the reference predictive distribution is obtained using a reference prior for all the parameters involved in the model so that, as he requires, the question of nuisance parameters is then avoided.

I am grateful to Professor Good for his comments. As A. P. Dempster once remarked, "In the area of statistical inference, there must be little that anyone has thought about that Dr Good has not written about, to the point that a computerized information retrieval system would be very helpful to scholars in the area."

I am aware of Professor Good's interest in multinomial problems, which also do intrigue me. I hope to be able in the near future to devote some time to study them from the perspective of this paper.

Professor Hartigan is certainly right in demanding more careful attention to mathematical detail if the procedure is to be systematically used. I insist, however, that this was premature before we could understand what reference posteriors really meant.

I believe that pathological cases as those mentioned by Professor Hartigan may be avoided with some mild regularity conditions for the class of admissible priors. Two such reasonable conditions are that (i) the joint measure $p(z, \theta)$ should be absolutely continuous with respect to the product measure $p(z)\,p(\theta)$ (see Osteyee and Good, 1974, p. 32) and (ii) the priors $p(\theta)$ should be strictly positive. It is clear that none of Professor Hartigan's examples meets these conditions.

I am grateful to Professor Press by his encouraging comments and by bringing to my attention an interesting paper which I had overlooked.

I have not fully investigated the implications of the approximations used but I suspect that one is only imposing mild regularity conditions to the class of admissible priors such as those mentioned above in reply to Professor Hartigan.

Dr Skene wonders what is the use of a marginal distribution $\int \pi(\theta, \omega \mid x)\, d\theta$ from a joint reference posterior $\pi(\theta, \omega \mid x)$. None, I would say, unless the operational prior which produces the joint reference posterior $\pi(\theta, \omega \mid x)$ happens to coincide with the operational prior which produces the reference posterior for ω.

Inference about the two variance components in the random effects model is an important problem where no generally accepted solution exists. I would be glad if Dr Skene could devote some of his time to produce and analyse the relevant reference posterior distributions.

I would like to thank all the discussants for the stimulus they have provided in making me think about the issues raised in the paper. The best way to understand an argument is, possibly, to be obliged to defend it.

References in the Discussion

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, Mass.: Addison-Wesley.

Fraser, D. A. S. (1972). Events, information processing and the structured model. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 32–55. Toronto: Holt, Reinhart and Winston.

—— (1974). Comparison of inference philosophies. In *Information, Inference and Decision* (G. Menges, ed.), pp. 77–98. Dordrecht: Reidel.

Geisser, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference* (V. P. Godambe and A. D. Sprott, eds), pp. 456–469. Toronto: Holt, Reinhart and Winston.

—— (1977). Discussion of the paper by G. N. Wilkinson. *J. R. Statist. Soc.* B, **39**, 155–156.

Good, I. J. (1952). Rational decisions. *J. R. Statist. Soc.* B, **14**, 107–114.

—— (1960). Discussion of a paper by E. M. L. Beale. *J. R. Statist. Soc.* B, **22**, 79–82.

—— (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press.

—— (1967). A Bayesian significance test for multinomial distributions (with Discussion). *J. R. Statist. Soc.* B, **29**, 399–431.

—— (1968). Utility of a distribution. *Nature*, **219**, 1392.

—— (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. of Statist.*, **4**, 1159–1189.

—— (1979). Turing's statistical work in World War II. *Biometrika*, **66** (in press).

Good, I. J. and Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Ass.*, **69**, 711–720.

Hill, B. M. (1976). Exact and approximate solutions for inference about variance components and multivariate inadmissibility. Technical Report AFFDL–TR–75–134. Air Force Flight Dynamics Laboratory, Wright–Patterson AFB, Ohio.

Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognisers. *IEEE Trans. Inf. Theory*, **14**, 55–63.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630.

—— (1978). Marginalization and prior probabilities. In *Studies of Bayesian Statistics* (A. Zellner, ed.). Amsterdam: North Holland.

Kullback, S. (1959). *Information Theory and Statistics.* New York: Wiley.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

Lindley, D. V. (1978). The Bayesian approach. *Scand. J. Statist.*, **5**, 1–26.

Murray, G. D. (1977). A note on the estimation of probability density functions. *Biometrika*, **64**, 150–152.

Osteyee, D. B. and Good, I. J. (1974). *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection.* Berlin: Springer-Verlag. (Lecture Notes in Mathematics, No. 376.)

Press, S. J. and Zellner, A. (1978). Posterior distribution for the multiple correlation coefficient with fixed regressors. *J. Econometrics*, **8**, 307–321.

Renyi, A. (1962/70). *Probability Theory.* Amsterdam: North Holland.

Thatcher, A. R. (1964). Relationships between Bayesian and confidence limits for predictions. *J. R. Statist. Soc.* B, **26**, 176–210.