Marginalization Paradoxes in Bayesian and Structural Inference

By A. P. DAWID, M. STONE and J. V. ZIDEK

University College London

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, February 14th, 1973, Professor J. DURBIN in the Chair]

SUMMARY

We describe a range of routine statistical problems in which marginal posterior distributions derived from improper prior measures are found to have an unBayesian property—one that could not occur if proper prior measures were employed. This paradoxical possibility is shown to have several facets that can be successfully analysed in the framework of a general group structure. The results cast a shadow on the uncritical use of improper prior measures.

A separate examination of a particular application of Fraser's structural theory shows that it is intrinsically paradoxical under marginalization.

Keywords: IMPROPER PRIORS; MARGINAL POSTERIORS; REDUCIBILITY; MARGINALIZATION PARADOX; GROUP ANALYSIS; ORBITS; RIGHT-INVARIANT PRIOR; MAXIMAL INVARIANT; SUBGROUP; STRUCTURAL INFERENCE INCONSISTENCIES

0. INTRODUCTION

A PROBLEM that has occupied the attention of many statisticians, particularly in recent years, has been the search for a mathematical expression of the state of ignorance about a parameter in a statistical model. Within a Bayesian framework, this ignorance is often supposed to be expressible by a particular prior distribution, which in almost all cases of interest is improper, so that it gives infinite "probability" to the whole parameter space. Many authors have tried to develop criteria for constructing such ignorance priors. One approach, exemplified by Novick (1969), is to use a prior that may be considered a limiting case of proper prior distributions that are conjugate to the family of sampling distributions (Raiffa and Schlaifer, 1961). The most widespread alternative approach concentrates on properties of *invariance* that should be satisfied by any satisfactory method of assigning ignorance prior distributions to statistical models (Jeffreys, 1961; Hartigan, 1964; Villegas, 1971). Although no fully acceptable theory has been developed, improper priors are in widespread current use among Bayesian statisticians, the recent book by Zellner (1971) containing many interesting examples.

It has usually been implicitly assumed that, for inferential purposes, improper prior distributions behave like proper ones, and they have been used with few qualms. However, some rather puzzling features have already been discovered. For the general multivariate normal model, Geisser and Cornfield (1963) show that there is no prior distribution for which the posterior distribution of the pivotal Hotelling's T^2 is the same as its sampling distribution, while at the same time the posterior distribution of Student's t, based on just one component of the data, is the same as its sampling distribution. Both of these requirements seem to express prior ignorance,

and both have been used to construct fiducial distributions, but they are inconsistent. This inconsistency was taken by Wilkinson (1971) as an argument against Bayesian inference in general! Yet another inconsistency, effectively the paradox of Example 4b below, is implicit in the work of Geisser (1965) although he does not appear to have noticed its significance.

It is not only Bayesians who have been interested in the expression of ignorance by means of invariance. This concept is useful in fiducial theory, while Fraser's (1968) theory of structural inference makes it almost axiomatic. It fact there are strong contacts between structural inference and Bayesian inference using invariant prior distributions (Fraser, 1961; Bondar, 1972). Another application of the concept is in the elimination of nuisance parameters from a likelihood function (Kalbfleisch and Sprott, 1970).

In Section 1 we demonstrate, by example, a paradox, which we call the marginalization paradox, that can arise from the use of improper prior distributions. Sections 2 and 3 develop in detail a group-theoretical analysis of this paradox, while Section 4, which is almost self-contained, uncovers some related inconsistencies of structural inference. If the lesson of these examples and their analysis is taken to heart, it may be that more statisticians will be guided by the philosophy of Lindley and Smith (1972) and turn their attentions to the characterization of prior knowledge, rather than prior ignorance.

1. THE MARGINALIZATION PARADOX

The paradox of central concern in this paper was first described by Stone and Dawid (1972). Before proceeding to a more general and extensive analysis than was undertaken in that paper, it will be useful to present further examples of statistical interest.

In order to clarify the hierarchy of paradox possible, our examples will be presented as involving two Bayesians, namely, B_1, who believes in using the whole data in any analysis, and B_2, who always arrives late on the scene of inference and who is ready to exploit any features that lead to a simplified analysis.

Example 1. *The change-point problem.* Suppose that

(i) observations have been taken of n successive, independent, exponentially distributed intervals $x_1, \ldots, x_n$;
(ii) it is known that the first ζ of these intervals have expectations $1/\eta$ and the remaining $n-\zeta$ have expectation $1/(c\eta)$;
(iii) c is known and $c \neq 1$, ζ is known only to take a value in $\{1, 2, \ldots, n-1\}$, while η is not known.

The probability density function of $x = (x_1, \ldots, x_n)$ equals

$$c^{n-\zeta}\eta^n \exp\left\{-\eta\left(\sum_1^{\zeta} x_i + c\sum_{\zeta+1}^{n} x_i\right)\right\}.$$

Our first Bayesian, B_1, chooses an improper prior distribution for $\theta = (\eta, \zeta)$ with measure element $\pi(d\theta) = \pi(\zeta)\,d\eta$ where $\pi(1) + \ldots + \pi(n-1) = 1$. Integrating out η gives the posterior probability distribution of ζ

$$\pi(\zeta \mid x) \propto \pi(\zeta)\left(\sum_1^{\zeta} z_i + c\sum_{\zeta+1}^{n} z_i\right)^{-(n+1)} c^{-\zeta}, \tag{1.1}$$

where $z_i = x_i/x_1$ $(i = 1, \ldots, n)$. Then B_2 arrives and notices

(i) the posterior distribution (1.1) is a function of $z = (z_1, \ldots, z_n)$ only;

(ii) the probability density function for z is a function of ζ only, in fact,

$$f(z \mid \eta, \zeta) = f(z \mid \zeta) \propto \left(\sum_1^{\zeta} z_i + c \sum_{\zeta+1}^{n} z_i \right)^{-n} c^{-\zeta}; \tag{1.2}$$

(iii) the right-hand side of (1.2) is not a factor of the right-hand side of (1.1), that is, for no choice of function $\pi^*(\zeta)$ is $\pi(\zeta \mid x)$ proportional to $f(z \mid \zeta)\pi^*(\zeta)$.

So B_2 is unable to reproduce (1.1) by any use of Bayes's theorem in conjunction with (1.2). That is, B_2's intervention has revealed the paradoxical unBayesianity of B_1's posterior distribution for ζ. However, the paradox is here easily avoidable if B_1 changes $\pi(d\theta)$ from $\pi(\zeta)\, d\eta$ to $\pi(\zeta)\, d\eta/\eta$, in agreement with the usual prescription for a scale parameter.

Example 2. *Discrimination parameter for two populations.* With data $x = (u_1, u_2, s^2)$, which are independently distributed with $u_1 \sim N(\mu_1, \sigma^2)$, $u_2 \sim N(\mu_2, \sigma^2)$ and $s^2 \sim \sigma^2 \chi_\nu^2/\nu$, inference is required about the "discrimination parameter" $\zeta = (\mu_1 - \mu_2)/(\sigma\sqrt{2})$.

In the light of his experience in Example 1, B_1 confidently employs the widely recommended prior having measure element $d\mu_1\, d\mu_2\, d\sigma/\sigma$. He finds the posterior probability element for ζ to be given by

$$d\zeta \int_0^{\infty} \omega^{\nu-1} \exp\left[-\tfrac{1}{2}\{\nu\omega^2 + (z\omega - \zeta)^2\}\right] d\omega, \tag{1.3}$$

where $z = (u_1 - u_2)/(s\sqrt{2})$. Unfortunately B_2 can again put his oar in by noticing that

(i) the posterior distribution for ζ is a function of z only;

(ii) the probability density function for z is a function of ζ only, in fact,

$$f(z \mid \mu_1, \mu_2, \sigma^2) = f(z \mid \zeta) \propto \int_0^{\infty} \omega^{\nu} \exp\left[-\tfrac{1}{2}\{\nu\omega^2 + (z\omega - \zeta)^2\}\right] d\omega; \tag{1.4}$$

(iii) the right-hand side of (1.4) is not a factor of the right-hand side of (1.3).

So, once more, B_2 is unable to derive B_1's posterior distribution by the use of Bayes's theorem in conjunction with what is, for him, the relevant likelihood function. In *this* example the paradox would not have arisen if B_1 had used the prior element $d\mu_1\, d\mu_2\, d\sigma/\sigma^2$ for which no recommendations appear to exist.

Example 3. *Correlation coefficient in the "progression model".* The data consist of n independent observations of a bivariate random variable (x_1, x_2) having a distribution given by

$$x_1 = \sigma_1 e_1, \quad x_2 = \gamma x_1 + \sigma_2 e_2, \tag{1.5}$$

where e_1, e_2 are independent $N(0, 1)$ random variables and $\theta = (\gamma, \sigma_1, \sigma_2)$ is unknown with $-\infty < \gamma < \infty$, $\sigma_1 > 0$, $\sigma_2 > 0$.

Inference is required about ζ, the correlation coefficient of x_1 and x_2, given by

$$\zeta = \gamma\sigma_1/(\gamma^2\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}. \tag{1.6}$$

Had it not been for his experience with Example 2, B_1 would have adopted without hesitation the prior measure element, having a "recommended" appearance,

$$\pi(d\theta) = d\gamma \frac{d\sigma_1}{\sigma_1} \frac{d\sigma_2}{\sigma_2}. \tag{1.7}$$

He is encouraged to adopt (1.7) by a preliminary analysis showing that (1.5) is a re-parametrization of a zero mean version of the "progression model" of Fraser (1968, p. 139 *et seq.*) and that (1.7) would reproduce in Bayesian terms the structural analysis of that model. That is, the use of (1.7) would give a posterior distribution for θ identical with Fraser's structural distribution. This identity inspires confidence, since, while Fraser's theory is somewhat controversial at the initial axiom level, it is not known to lead to any paradoxical behaviour.

Writing

$$S_{11} = \sum x_{i1}^2, \quad S_{12} = \sum x_{i1} x_{i2}, \quad S_{22} = \sum x_{i2}^2,$$

the posterior distribution for θ has kernel

$$(\sigma_1 \sigma_2)^{-(n+1)} \exp\left[-\frac{1}{2(1-\zeta^2)}\left\{\frac{S_{11}}{\sigma_1^2} - \frac{2\zeta S_{12}}{\sigma_1(\gamma^2\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}} + \frac{S_{22}}{\gamma^2\sigma_1^2+\sigma_2^2}\right\}\right] d\gamma\, d\sigma_1\, d\sigma_2$$

$$= (\sigma_1 \sigma_2)^{-(n+1)} \exp\left[-\frac{1}{2(1-\zeta^2)}\left\{\frac{S_{11}}{\sigma_1^2} - \frac{2\zeta(1-\zeta^2)^{\frac{1}{2}} S_{12}}{\sigma_1 \sigma_2} + \frac{S_{22}(1-\zeta^2)}{\sigma_2^2}\right\}\right] d\gamma\, d\sigma_1\, d\sigma_2. \tag{1.8}$$

The Jacobian of the transformation $\theta \to (\zeta, \sigma_1, \sigma_2)$ is $\sigma_1(1-\zeta^2)^{\frac{1}{2}}/\sigma_2$, whence the posterior distribution of ζ has kernel

$$d\zeta(1-\zeta^2)^{-\frac{1}{2}} \int\int \sigma_1^{-(n+2)} \sigma_2^{-n} \exp\left[-\frac{1}{2(1-\zeta^2)}\left\{\frac{S_{11}}{\sigma_1^2} - \frac{2\zeta(1-\zeta^2)^{\frac{1}{2}} S_{12}}{\sigma_1 \sigma_2} + \frac{S_{22}(1-\zeta^2)}{\sigma_2^2}\right\}\right] d\sigma_1\, d\sigma_2.$$

The substitutions $v = S_{11}/\{\sigma_1^2(1-\zeta^2)\}$, $\psi = (1-\zeta^2)^{\frac{1}{2}}(\sigma_1 S_{22}^{\frac{1}{2}})/(\sigma_2 S_{11}^{\frac{1}{2}})$ show that this is proportional to

$$d\zeta(1-\zeta^2)^{\frac{1}{2}(n-2)} \int_0^\infty \psi^{n-2}\, d\psi \int_0^\infty v^{n-1} \exp\{-\tfrac{1}{2}(1-2z\zeta\psi+\psi^2)\, v\}\, dv$$

$$\propto d\zeta(1-\zeta^2)^{\frac{1}{2}(n-2)} \int_0^\infty \frac{d\psi}{\psi^2} (\psi - 2z\zeta + \psi^{-1})^{-n}, \tag{1.9}$$

where $z = S_{12}/(S_{11} S_{22})^{\frac{1}{2}}$, the sample correlation coefficient of x_1 and x_2.

B_2 notes that (1.9) is a function of z alone and, yet again, the probability density function for z depends only on ζ. In fact

$$f(z|\gamma, \sigma_1, \sigma_2) = \text{const} \times (1-z^2)^{\frac{1}{2}(n-3)} (1-\zeta^2)^{\frac{1}{2}n} \int_{-\infty}^\infty \frac{dy}{(\cosh y - z\zeta)^n} \tag{1.10}$$

(Zellner, 1971, p. 390) and substituting $\psi = e^y$ yields

$$\text{const} \times (1-z^2)^{\frac{1}{2}(n-3)} (1-\zeta^2)^{\frac{1}{2}n} \int_0^\infty \frac{d\psi}{\psi} (\psi - 2z\zeta + \psi^{-1})^{-n}.$$

Comparison of this and (1.9) reveals that (1.7) would, after all, lead to paradox.

B_1 then observes that the progression model is equivalent, from a non-structural point of view, to the statement that (x_1, x_2) is normally distributed with zero mean and covariance matrix $\boldsymbol{\Sigma}$. Moreover, in this formulation, there is the widely recommended and utilized prior measure element $d\boldsymbol{\Sigma}/|\boldsymbol{\Sigma}|^{\frac{3}{2}}$, the special case of the p-dimensional $d\boldsymbol{\Sigma}/|\boldsymbol{\Sigma}|^{\frac{1}{2}(p+1)}$. This choice is equivalent to $d\gamma\, d\sigma_1\, d\sigma_2/\sigma_2^2$ or

$$(1-\zeta^2)^{-\frac{1}{2}}\, d\zeta (d\sigma_1/\sigma_1)(d\sigma_2/\sigma_2)$$

in the alternative parametrizations. B_1 is relieved to find that, for this prior, there is no paradox for inference about ζ.

Example 4a. *Coefficients of variation.* The data $(x_{11}, \ldots, x_{1n})$, $(x_{21}, \ldots, x_{2n})$ are two independent samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively. While $\theta = (\mu_1, \mu_2, \sigma)$, interest is centred on the two parameters $\zeta_1 = \mu_1/\sigma$ and $\zeta_2 = \mu_2/\sigma$. In the light of his experience with Examples 1, 2 and 3, B_1 decides to consider a prior measure element of the form

$$\pi(d\theta) = \sigma^p \, d\mu_1 \, d\mu_2 \, d\sigma \tag{1.11}$$

reserving his commitment to it and deferring the choice of p until he has examined the implications of the choice as far as possible paradoxical interaction with B_2 is concerned. With (1.11), B_1's posterior distribution for (ζ_1, ζ_2) has kernel

$$\int_0^\infty \omega^{2n-4-p} \exp\left[-\tfrac{1}{2}\{\omega^2 + n(z_1\omega - \zeta_1)^2 + n(z_2\omega - \zeta_2)^2\}\right] d\omega, \tag{1.12}$$

where

$$z_i = \bar{x}_i/s \quad \text{with} \quad \bar{x}_i = \sum_j x_{ij}/n, \quad s^2 = \sum\sum (x_{ij} - \bar{x}_i)^2.$$

Noting, before B_2 arrives, that (1.12) depends only on (z_1, z_2) and that the probability density of (z_1, z_2) depends only on (ζ_1, ζ_2) and is proportional to

$$\int_0^\infty \omega^{2n-1} \exp\left[-\tfrac{1}{2}\{\omega^2 + n(z_1\omega - \zeta_1)^2 + n(z_2\omega - \zeta_2)^2\}\right] d\omega, \tag{1.13}$$

B_1 decides that, if he is to avoid paradoxical conflict with B_2, he should take $p = -3$. For, with that choice, (1.13) will be a factor of (1.12).

However B_2 asserts his interest in ζ_1 alone. He finds that B_1's posterior density for ζ_1, with the choice $p = -3$ in (1.11), has kernel

$$d\zeta_1 \int_0^\infty \omega^{2n-1} \exp\left[-\tfrac{1}{2}\{\omega^2 + n(z_1\omega - \zeta_1)^2\}\right] d\omega \tag{1.14}$$

which involves only z_1; while the probability density of z_1 depends only on ζ_1 and is proportional to

$$d\int_0^\infty \omega^{2n-2} \exp\left[-\tfrac{1}{2}\{\omega^2 + n(z_1\omega - \zeta_1)^2\}\right] d\omega. \tag{1.15}$$

Once more, B_2 cannot match B_1's inference about a parameter of interest using a combination of any prior with what appears to be the appropriate likelihood function, here (1.15). B_1 is rather mortified to find that if he had only known that B_2 was interested in ζ_1 alone, rather than (ζ_1, ζ_2), he would have been able to make a harmonizing choice of p, namely $p = -2$!

Example 4b. *Correlation coefficients among three variables.* The behaviour uncovered in Example 4a occurs also in the following example of wide statistical interest. The data consist of n independent observations of a trivariate normal random variable (x_1, x_2, x_3) having mean zero and unknown covariance matrix $\boldsymbol{\Sigma}$. The recommended prior element successfully employed in Example 3 becomes $d\boldsymbol{\Sigma}/|\boldsymbol{\Sigma}|^2$ for $p = 3$. A slight modification of Geisser's analysis (1965, p. 154) shows that the marginal posterior density of ζ, the correlation coefficient of x_1 and x_2, is proportional to

$$(1 - \zeta^2)^{\frac{1}{2}(n-4)} I_{n-1}(z\zeta), \tag{1.16}$$

where

$$I_\nu(\xi) = \int_0^\infty \frac{dy}{(\cosh y - \xi)^\nu}$$

and z is the sample correlation coefficient of x_1 and x_2.

This depends on z alone, while the sampling density of z, given by (1.10), is proportional to

$$(1-\zeta^2)^{\frac{1}{2}n} I_n(z\zeta) \tag{1.17}$$

which is not a factor of (1.16).

In fact B_1 finds that in the class of prior measures $d\boldsymbol{\Sigma}/|\boldsymbol{\Sigma}|^{\frac{1}{2}v}$ it is necessary to choose $v = 5$ in order to avoid a paradox for ζ. This he does, but B_2 then announces his interest in $\zeta_{12.3}$, the partial correlation coefficient of x_1 and x_2, given x_3. The posterior density for $\zeta_{12.3}$ implied by B_1's new prior is proportional to

$$(1-\zeta_{12.3}^2)^{\frac{1}{2}(n-2)} I_{n+1}(z_{12.3}\,\zeta_{12.3}), \tag{1.18}$$

where $z_{12.3}$ is the sample counterpart of $\zeta_{12.3}$, whereas the kernel of the sampling density of $z_{12.3}$, dependent only on $\zeta_{12.3}$, is

$$(1-\zeta_{12.3}^2)^{\frac{1}{2}(n-1)} I_{n-1}(z_{12.3}\,\zeta_{12.3}). \tag{1.19}$$

Thus B_1 again finds himself in a dilemma: there is no choice of v that simultaneously avoids paradox for both ζ and $\zeta_{12.3}$.

In the confusion into which they are thrown by these five examples, B_1 and B_2 are justified in believing that further analysis is needed.

Such confusion could not have arisen if B_1 had employed proper prior distributions integrating to unity. To see this, let us adopt a general framework. Data $x = (y, z)$ in space $Y \times Z$ has density element

$$f(x|\,\theta)\,dx = f(y, z|\,\eta, \zeta)\,dy\,dz$$

depending on some parameter $\theta = (\eta, \zeta)$, with the property that the density of z is a function $f(z|\,\zeta)$ of ζ only. B_1 uses some prior distribution for θ, represented by the measure element $\pi(d\theta) = \pi(d\eta, d\zeta)$, yielding a posterior probability element for ζ given by

$$\pi_1(d\zeta|\,x) = \int_\eta f(y, z|\,\eta, \zeta)\,\pi(d\eta, d\zeta) \Big/ \int f(x|\,\theta)\,\pi(d\theta) = a(z, \zeta, d\zeta), \tag{1.20}$$

say, by supposition. Whence

$$f(z|\,\zeta) \int_\eta f(y|\,z, \eta, \zeta)\,\pi(d\eta, d\zeta) = a(z, \zeta, d\zeta) \int f(x|\,\theta)\,\pi(d\theta). \tag{1.21}$$

If π is proper, that is, if $\int \pi(d\theta) = 1$, we may integrate both sides of (1.21) with respect to y to give

$$f(z|\,\zeta)\,\pi(d\zeta) = a(z, \zeta, d\zeta) \int f(z|\,\zeta)\,\pi(d\zeta), \tag{1.22}$$

where $\pi(d\zeta) = \int_\eta \pi(d\zeta, d\eta)$. From (1.20) and (1.22) we see that $\pi_1(d\zeta|\,x)$ will be identical with $\pi_2(d\zeta|\,z)$ given by

$$\pi_2(d\zeta|\,z) \propto f(z|\,\zeta)\,\pi(d\zeta) \tag{1.23}$$

showing that $\pi(d\zeta)$ is the compatible choice of prior element for B_2, as is to be expected.

In the next section, we introduce and employ the tools of group theory in further analysis of the paradox.

2. GROUP ANALYSIS

2.1 *Data Space, Parameter Space and Constancy*

Suppose we are given data x distributed over data space X with probability distribution P_θ conditional on θ. P_θ is known only to lie in the class $\{P_\theta \mid \theta \in \Theta\}$ where Θ is the parameter space. Suppose

$$\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}. \tag{2.1}$$

Our model statistical problem is that of inference about some parameter ζ, a function $\zeta(\theta)$ of θ. We consider a (perhaps trivial) group G of one-to-one transformations of X onto itself, represented by $g: x \to g \circ x$ such that

(i) for each θ, x distributed as P_θ implies that $g \circ x$ is distributed as $P_{\bar{g} \circ \theta}$, say, where $\bar{g} \circ \theta \in \Theta$,
(ii) $\zeta(\bar{g} \circ \theta) \equiv \zeta(\theta)$.

It can be shown that $\bar{G} = \{\bar{g} \mid g \in G\}$ is a group of one-to-one transformations of Θ onto itself. Moreover $\bar{G}$ is homomorphic to G, that is, $\overline{(g^{-1})} = (\bar{g})^{-1}$ and $\overline{(gh)} = \bar{g}\bar{h}$. The *orbit under* G of the point x is the set

$$G \circ x = \{g \circ x \mid g \in G\}$$

and the orbit under $\bar{G}$ of the point θ is

$$\bar{G} \circ \theta = \{\bar{g} \circ \theta \mid \bar{g} \in \bar{G}\}.$$

The orbits partition their respective spaces. Clearly ζ is constant on each orbit under $\bar{G}$ in Θ.

In the subclass of problems to be considered, we shall require ζ to be a *maximal invariant* under $\bar{G}$, that is, ζ takes different values on any two orbits. In this case, we will say that the problem is *constant under* G.

From each orbit $G \circ x$, we select a representative element, $z(x)$ say. The parliament of these representatives is $\{z(x) \mid x \in X\} = Z$, say. Then $z(g \circ x) = z(x)$, $g \in G$, $x \in X$. We suppose that the group G is *exact*, that is, for any $x \in X$, $g_1 \circ x = g_2 \circ x \Rightarrow g_1 = g_2$. It follows that for each x there is a unique member of G, $y(x)$ say, such that

$$x \equiv y(x) \circ z(x), \quad y(g \circ x) \equiv g y(x). \tag{2.2}$$

This allows X to be identified with $G \times Z$ by $x \sim (y(x), z(x))$. For clarity in the sequel, G is denoted by Y when used in this identification, which is then

$$X = Y \times Z. \tag{2.3}$$

If $\bar{G}$ is also exact, we may likewise obtain $\Theta = \mathcal{Y} \times \mathcal{Z}$ where $\mathcal{Y} = \bar{G}$ and $\mathcal{Z}$ ($\subset \Theta$) is the parliament of representatives of the orbits under $\bar{G}$ in Θ. Because ζ is the maximal invariant under $\bar{G}$, labelling the orbits $\{\bar{G} \circ \theta\}$, we may, for convenience, identify $\zeta(\theta)$ with the representative of $\bar{G} \circ \theta$.

The problem can now be restated in terms of the data (y, z), distributed over $Y \times Z$ according to P_θ conditional on the parameter $\theta = (\eta, \zeta) \in \mathcal{Y} \times \mathcal{Z}$.

In the sequel, only locally compact topological groups are considered, a restriction that is of no practical importance. We employ some familiar results concerning such

transformation groups (Nachbin, 1965). If S is any such group then it will have a *left-invariant measure* μ_S:

$$\mu_S(sA) \equiv \mu_S(A), \quad s \in S, \quad A \subset S$$

and a *right-invariant measure* ν_S:

$$\nu_S(As) \equiv \nu_S(A), \quad s \in S, \quad A \subset S.$$

These measures are unique up to multiplicative constants. We can and do choose $\nu_S(A) = \mu_S(A^{-1})$ where $A^{-1} = \{a^{-1} | a \in A\}$. There is a *modular function* Δ_S with the property

$$\mu_S(As) \equiv \Delta_S(s)\mu_S(A).$$

Δ_S is a *morphism*, that is, Δ_S is positive, continuous and

$$\Delta_S(s_1 s_2^{-1}) \equiv \Delta_S(s_1)\{\Delta_S(s_2)\}^{-1}.$$

It can be shown that

$$\mu_S(ds) \equiv \nu_S(ds^{-1}) \equiv \Delta_S(s)\nu_S(ds).$$

Any measure μ^* on Θ satisfying

$$\frac{\mu^*(\bar{g} \circ A)}{\mu^*(\bar{g} \circ B)} \equiv \frac{\mu^*(A)}{\mu^*(B)} \tag{2.4}$$

$\bar{g} \in \bar{G}$, $A \subset \Theta$, $B \subset \Theta$ is called *relatively invariant* with respect to $\bar{G}$. Such measures are, alternatively, characterized by the equivalent property

$$\mu^*(\bar{g} \circ A) \equiv \xi(\bar{g})\mu^*(A), \tag{2.5}$$

where ξ is a morphism on $\bar{G}$.

We suppose that P_θ has density element

$$f(y, z | \eta, \zeta)\mu_G(dy)\,dz \tag{2.6}$$

where dz denotes a fixed general measure element that need not be specified for our analysis.

Problems that are constant under G are then those for which

$$f(y, z | \eta, \zeta) \equiv f(gy, z | \bar{g}\eta, \zeta). \tag{2.7}$$

An important consequence is expressed in the following lemma.

Lemma 2.1. If (2.7) holds, the distribution of z depends only on ζ and has density element

$$f(z | \zeta)\,dz \propto \left\{ \int f(e, z | \eta, \zeta)\nu_{\bar{G}}(d\eta) \right\} dz, \tag{2.8}$$

where e is the identity element of Y.

Proof.

$$f(z | \eta, \zeta) = \int f(y, z | \eta, \zeta)\mu_G(dy) = \int f(e, z | \bar{y}^{-1}\eta, \zeta)\mu_G(dy)$$

$$= \int f(e, z | \eta', \zeta)m(d\eta'),$$

where m is the measure induced on $\bar{y}^{-1}\eta$ by μ_G. It may be verified that m is right-invariant, so that $m \propto \nu_{\bar{G}}$ and (2.8) follows.

It remains to consider the choice of prior distribution for θ. Many statisticians would think it reasonable to restrict the choice to a prior from the class for which any posterior inferences about θ would be invariant under the concerted action of the groups G and $\bar{G}$; that is, expressing this condition in terms of the posterior probability distribution,

$$\pi(\theta \in A \,|\, x) \equiv \pi(\theta \in \bar{g} \circ A \,|\, g \circ x).$$

An adaptation of Stone (1970) shows that, under weak conditions, this implies that the prior for θ must satisfy

$$\pi(\bar{g} \circ A) \equiv \alpha(\bar{g})\, \pi(A), \tag{2.9}$$

where α is a morphism on $\bar{G}$; that is, the prior must be relatively invariant under $\bar{G}$. A further simple argument, based on the essential uniqueness of the left invariant measure on $\bar{G}$, shows that (2.9) implies that the prior measure element for (η, ζ) has the product form

$$\pi(d\eta, d\zeta) \propto \xi(\eta)\, \nu_{\bar{G}}(d\eta)\, d\zeta \tag{2.10}$$

where $\xi(\eta) \equiv \Delta_{\bar{G}}(\eta)\, \alpha(\eta)$ and $d\zeta$ denotes an arbitrary measure element. Conversely if (2.10) obtains for some morphism ξ on $\mathcal{G}$ then so does (2.9) with

$$\alpha(\bar{g}) \equiv \xi(\bar{g}) \{\Delta_{\bar{G}}(\bar{g})\}^{-1}.$$

Note that ξ, as a morphism, must be identically 1 if it is a constant.

2.2. *Examples*

Table 1 shows the specialization of the general model for Examples 1–3 of Section 1. (The data of Example 3 have been reduced to the sufficient statistics.)

TABLE 1

Components of the general model in three examples

General	*Example* 1	*Example* 2	*Example* 3
Data	$(x_1, \ldots, x_n)$	(u_1, u_2, s^2)	(S_{11}, S_{22}, S_{12})
θ	(η, ζ)	(μ_1, μ_2, σ^2)	$(\gamma, \sigma_1, \sigma_2)$
G	$\{[a] \mid a > 0\}$	$\{[a, b] \mid a > 0,\ -\infty < b < \infty\}$	$\{[a, b] \mid a > 0,\ b > 0\}$
$g \circ x$	$(ax_1, \ldots, ax_n)$	$(au_1 + b,\ au_2 + b,\ a^2 s^2)$	$(a^2 S_{11},\ b^2 S_{22},\ ab\, S_{12})$
g^{-1}	$[a]^{-1} = [a^{-1}]$	$[a, b]^{-1} = [a^{-1}, -ba^{-1}]$	$[a, b]^{-1} = [a^{-1}, b^{-1}]$
$g_1 g_2$	$[a]\,[b] = [ab]$	$[a, b]\,[c, d] = [ac,\ ad + b]$	$[a, b]\,[c, d] = [ac,\ bd]$
$y(x)$	$[x_1]$	$[s, u_1]$	$[S_{11}^{\frac{1}{2}},\ S_{22}^{\frac{1}{2}}]$
$z(x)$	$(1, x_2/x_1, \ldots, x_n/x_1)$	$(0, (u_2 - u_1)/s, 1)$	$(1, 1, z)$
$\bar{G}$	$\{\{a\} \mid a > 0\}$	$\{\{a, b\} \mid a > 0,\ -\infty < b < \infty\}$	$\{\{a, b\} \mid a > 0,\ b > 0\}$
$\bar{g} \circ \theta$	$(a\eta, \zeta)$	$(a\mu_1 + b,\ a\mu_2 + b,\ a^2 \sigma^2)$	$(\gamma b/a,\ a\sigma_1,\ b\sigma_2)$
$g \to \bar{g}$	$[a] \to \{a^{-1}\}$	$[a, b] \to \{a, b\}$	$[a, b] \to \{a, b\}$
$\zeta(\theta)$	$(1, \zeta)$	$(0, (\mu_2 - \mu_1)/\sigma, 1)$	$(\gamma\sigma_1/\sigma_2, 1, 1)$
$\eta(\theta)$	$\{\eta\}$	$\{\mu_1, \sigma\}$	$\{\sigma_1, \sigma_2\}$
$\mu_G(dy)$	dx_1/x_1	$du_1\, ds/s^2$	$dS_{11}\, dS_{22}/(S_{11}\, S_{22})$
$\nu_{\bar{G}}(d\eta)$	$d\eta/\eta$	$d\mu_1\, d\sigma/\sigma$	$d\sigma_1\, d\sigma_2/(\sigma_1\, \sigma_2)$
$\Delta_{\bar{G}}(\eta)$	1	σ^{-1}	1

2.3. *The Marginalization Paradox and its Avoidance*

Theorem 2.1. If the prior distribution is given by (2.10) then the posterior (marginal) distribution of ζ depends only on z and has density element

$$\pi(d\zeta \mid x) \propto \left\{ \int f(e, z \mid \eta, \zeta)\, \xi(\eta)\, \nu_{\bar{G}}(d\eta) \right\} d\zeta. \tag{2.11}$$

Proof.

$$\begin{aligned} \pi(d\zeta \mid x) &\propto \left\{ \int f(y, z \mid \eta, \zeta)\, \xi(\eta)\, \nu_{\bar{G}}(d\eta) \right\} d\zeta \\ &= \left\{ \int f(e, z \mid \bar{y}^{-1}\eta, \zeta)\, \xi(\eta)\, \nu_{\bar{G}}(d\eta) \right\} d\zeta \\ &\propto \left\{ \int f(e, z \mid \eta', \zeta)\, \xi(\eta')\, \nu_{\bar{G}}(d\eta') \right\} d\zeta \end{aligned}$$

by the properties of ξ and $\nu_{\bar{G}}$.

Comparing Lemma 2.1 and Theorem 2.1, we see that the choice of (2.10) with $\xi \not\equiv 1$ may well give us the marginalization paradox in the form: *Although z is the only aspect of the data needed to determine the posterior density of* ζ, *that density does not contain the probability density of z* (*dependent only on* ζ) *as a factor.* A Bayesian working with z alone could not match the consequences of (2.10).

The paradox does not arise if (i) $\xi \equiv 1$ or (ii) $f(y, z \mid \eta, \zeta) = f(y \mid z; \eta) f(z \mid \zeta)$ in which case ξ is not even required to be a morphism. (An example of (ii) occurs in estimating a principal sub-matrix of the unknown parameter matrix $\mathbf{\Sigma}$ of a Wishart variable $\mathbf{S} \sim W(\mathbf{\Sigma}, \nu, p)$. See Appendix 1(i).)

In most cases that arise in practice, such as Examples 1–3, setting $\xi \equiv 1$ is the only way to avoid the paradox, an indicative conclusion that may be stated: *Usually, use of the prior element* (2.10) *will lead to a marginalization paradox unless* $\xi(\eta) \equiv 1$ (*equivalent to* $\alpha(\bar{g}) \equiv \{\Delta_{\bar{G}}(\bar{g})\}^{-1}$ *in* (2.9)).

The condition $\xi(\eta) \equiv 1$ means that the prior for θ is given by the product of an arbitrary measure for ζ and right-invariant measure for η in $\bar{G}$. Hence our conclusion argues for the "rightness" of the choice of right-invariant measure from the class of relatively invariant measures. The next section shows how "two rights can make a wrong".

2.4. *The Group as Subgroup: Paradox Lost and Paradox Regained*

In many common problems having the structure of Section 2.1, there will be groups T, $\bar{T}$ of transformations on X, Θ, respectively, such that $\bar{T}$ is homomorphic to T, and

(i) $\bar{G}$ is a proper subgroup of $\bar{T}$.

(ii) The distributions are invariant under T and $\bar{T}$, that is, if x is distributed as P_θ, and $t \in T$, then $t \circ x$ is distributed as $P_{\bar{t} \circ \theta}$.

However, the problem involving ζ will not be constant under T since $\zeta(\bar{t} \circ \theta)$ and $\zeta(\theta)$ will not be equal for all $\bar{t} \in \bar{T}$, $\theta \in \Theta$.

There are two cases of particular interest.

Case 1. The group $\bar{T}$ is in one-to-one correspondence with Θ. Such a case occurs in Example 2, where we have $t = [a, b_1, b_2]$, $\bar{t} = \{a, b_1, b_2\}$ $(a > 0)$ with

$$t \circ (u_1, u_2, s^2) = (au_1 + b_1, au_2 + b_2, a^2 s^2)$$

and $\bar{t} \circ (\mu_1, \mu_2, \sigma^2) = (a\mu_1 + b_1, a\mu_2 + b_2, a^2\sigma^2)$. We can identify (μ_1, μ_2, σ^2) with $\bar{t} = \{\sigma, \mu_1, \mu_2\}$ so that $(\mu_1, \mu_2, \sigma^2) = \bar{t} \circ (0, 0, 1)$. Note that $\bar{G}$ is the subgroup of $\bar{T}$ obtained by the restriction $b_1 = b_2$.

In a case such as this, the only prior distributions for θ which lead to posterior distributions invariant under T and $\bar{T}$ are the relatively invariant measures on Θ considered as the group $\bar{T}$. Particular interest attaches to the choice of right-invariant measure on Θ considered as the group $\bar{T}$ (see Section 4).

Case 2. Suppose that $\bar{T}$ is not in one-to-one correspondence with Θ, but T and $\bar{T}$ are exact on X and Θ respectively. If $z^*(x)$, $\zeta^*(\theta)$ denote orbit labels of X, Θ under the action of T and $\bar{T}$, then $\zeta^*(\theta)$ is a function of $\zeta(\theta)$. This case arises in Example 4a with $\zeta = (\zeta_1, \zeta_2)$ and $\zeta^* = \zeta_1$. As in Section 2.1, the statistician may be interested in posterior distributions which are invariant under T and $\bar{T}$, implying a relatively invariant prior satisfying

$$\pi(\bar{t} \circ A) = \alpha(\bar{t})\,\pi(A), \quad \bar{t} \in \bar{T}, \quad A \subset \Theta \tag{2.12}$$

with α a morphism on $\bar{T}$.

The analysis of Sections 2.1–2.3 may be applied to the problem of finding the marginal distribution of ζ^*: we merely replace $G, \bar{G}$ by $T, \bar{T}$. Then, as the results of those sections indicate, a marginalization paradox will usually arise for ζ^* and z^*, unless $\alpha(\bar{t}) \equiv \{\Delta_{\bar{T}}(\bar{t})\}^{-1}$.

Note that this choice of α is just that which arises in Case 1 from the choice of a right-invariant prior on Θ.

If (2.12) holds for all $\bar{t} \in \bar{T}$, it holds *a fortiori* for all $\bar{g} \in \bar{G}$. Comparing with (2.9) and (2.10), we find $\xi(\bar{g}) \equiv \Delta_{\bar{G}}(\bar{g})\,\alpha(\bar{g}), (\bar{g} \in \bar{G})$. In particular, for the choice

$$\alpha(\bar{t}) \equiv \{\Delta_{\bar{T}}(\bar{t})\}^{-1},$$

which is of special importance for both Case 1 and Case 2 above, we get

$$\xi(\bar{g}) = \Delta_{\bar{G}}(\bar{g})\{\Delta_{\bar{T}}(\bar{g})\}^{-1}.$$

It is easy to find examples in which $\Delta_{\bar{G}}(\bar{g})\{\Delta_{\bar{T}}(\bar{g})\}^{-1}$ is *not* identically unity for $\bar{g} \in \bar{G}$, for instance in Examples 2 and 4a. When this is the case, the above choice of α will lead, usually, to a paradox for the marginal distribution of ζ. Hence we will be in the following situation:

Case 1. The use of right-invariant prior distribution on Θ leads to a marginalization paradox for ζ. (As in Example 2.)

Case 2. The use of the only relatively invariant prior which avoids a marginalization paradox for ζ^* entails such a paradox for ζ. (As in Example 4a.)

Case 2 is particularly worrying, since it means that there is *no* prior which is relatively invariant under $\bar{T}$ and does not exhibit paradoxical behaviour.

As a slight extension of the above analysis, we can relax the requirement that T and $\bar{T}$ be exact, and consider two subgroups G_1 and G_2 of T, leaving the problem constant for inference about ζ_1, ζ_2 respectively. Then there will usually be no prior, relatively invariant under $\bar{T}$, which simultaneously avoids a paradox for both ζ_1 and ζ_2, unless there is a morphism α on $\bar{T}$ which is equal to $\Delta_{\bar{G}_1}^{-1}$ on $\bar{G}_1$ and to $\Delta_{\bar{G}_2}^{-1}$ on $\bar{G}_2$.

The structure of Example 4b has special features which are considered in detail in Appendix 1.

3. RESTRICTED DATA PROBLEMS

In this section we show that the paradox may arise even for problems where there is no group structure, that is, where there is no non-trivial group of transformations under which the problem of inference about ζ is constant (see Section 2.1).

In the first example below, there is no group structure because B_1 has, contrary to his original principles, decided to eliminate some nuisance parameters by restricting the data on which he will base his inference. In the second example, the data are already in reduced form when B_1 receives them from a computer.

3.1. *Examples*

Example 5. *Coefficients of variation.* The data (x_1, x_2), say, have a probability density, dependent on $\theta = (\zeta, \xi)$, proportional to

$$\int_0^\infty t^{2n-1} \exp\left[-\tfrac{1}{2}\{t^2 + n(x_1 t - \zeta)^2 + n(x_2 t - \xi)^2\}\right] dt. \tag{3.1}$$

As (1.13) shows, (x_1, x_2) has the same distribution as $(\bar{u}/s, \bar{v}/s)$ where $\bar{u} = (u_1 + \ldots + u_n)/n$, $\bar{v} = (v_1 + \ldots + v_n)/n$, $s^2 = \sum(u_i - \bar{u})^2 + \sum(v_i - \bar{v})^2$ and $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ are independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, with the identification $\zeta = \mu_1/\sigma$, $\xi = \mu_2/\sigma$. The natural choice of prior measure element $d\zeta\, d\xi$ gives the posterior density element for ζ proportional to

$$d\zeta \int_0^\infty t^{2n-1} \exp\left[-\tfrac{1}{2}\{t^2 + n(x_1 t - \zeta)^2\}\right] dt, \tag{3.2}$$

dependent only on x_1. However x_1 has sampling density proportional to

$$\int_0^\infty t^{2n-2} \exp\left[-\tfrac{1}{2}\{t^2 + n(x_1 t - \zeta)^2\}\right] dt$$

which is not a factor of (3.2).

In Appendix 2, we show that there is, in fact, *no* prior non-degenerate measure for θ, giving a posterior distribution for ζ dependent only on x_1, that is free of this paradox.

Example 6. *Correlation matrices.* In this example, the raw data are $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, an independent sample from the p-dimensional multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ unknown. However, after processing by a computer, the only available data are the sample correlation matrix $\mathbf{R}$, obtained by standardizing the sample sum-of-squares-and-products matrix $\mathbf{S}$; that is, $r_{ij} = s_{ij}/(s_{ii}\, s_{jj})^{\frac{1}{2}}$, with

$$s_{ij} = \sum_{k=1}^{n} x_{ki}\, x_{kj} - \frac{1}{n}\left(\sum_{k=1}^{n} x_{ki}\right)\left(\sum_{k=1}^{n} x_{kj}\right).$$

Let $\boldsymbol{\Phi}$ denote the population correlation matrix, and $\boldsymbol{\Psi}$ the population standardized precision matrix. Thus $\phi_{ij} = \sigma_{ij}/(\sigma_{ii}\,\sigma_{jj})^{\frac{1}{2}}$, $\psi_{ij} = \sigma^{ij}/(\sigma^{ii}\,\sigma^{jj})^{\frac{1}{2}}$, where $\boldsymbol{\Sigma}^{-1} = (\sigma^{ij})$. (There is a one-to-one correspondence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, each being the standardized inverse of the other.) The sampling density of $\mathbf{R}$ is given by

$$f(\mathbf{R}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{R} \propto |\boldsymbol{\Psi}|^{\frac{1}{2}\nu}\, |\mathbf{R}|^{\frac{1}{2}(\nu - p - 1)}\, F_\nu(\boldsymbol{\Gamma})\, d\mathbf{R}, \tag{3.3}$$

where $\nu = n - 1$, $\gamma_{ij} = \psi_{ij} r_{ij}$ and

$$F_\nu(\boldsymbol{\Gamma}) = \int_0^\infty ds_1 \int_0^\infty ds_2 \ldots \int_0^\infty ds_p (s_1 s_2 \ldots s_p)^{\nu - 1} \exp\left(-\tfrac{1}{2}\mathbf{s}'\boldsymbol{\Gamma}\mathbf{s}\right)$$

(Fisher, 1962).

The distributions given by (3.3) depend only on Ψ (or equivalently Φ), but do not have any useful group-invariance properties.

With $\mathbf{R}$ as data, B_1 would like to make inferences about a principal $q \times q$ submatrix Φ_1 of Φ, and he decides to investigate the class of prior distributions with element $d\Phi/|\Phi|^{\frac{1}{2}v}$, equivalent to $|\Psi|^{\frac{1}{2}v-p-1}d\Psi$. (This distribution is suggested by decomposition of the element $d\mu\, d\Sigma/|\Sigma|^{\frac{1}{2}v}$; cf. Example 4b.)

He finds the posterior density to be given by

$$\pi(\Psi|\mathbf{R})\,d\Psi \propto |\Psi|^{\frac{1}{2}(\nu+v-2p-2)}F_\nu(\Gamma)\,d\Psi. \tag{3.4}$$

In general, the marginal distribution of Φ_1 obtained from (3.4) will depend on $\mathbf{R}$ as a whole. However, for the particular choice $v = p+1$, comparison of (3.3) and (3.4) shows that the roles of Ψ and $\mathbf{R}$ are interchanged. Thus the posterior distribution of Ψ is that of a Wishart variable $W(\mathbf{S}^{-1}, \nu, p)$ (Zellner, 1971, Appendix B) after standardization. Therefore, the posterior distribution of Φ is that of a standardized Inverted Wishart variable $IW(\mathbf{S}; \nu, p)$. But this inverted Wishart distribution is just that obtained for Σ from the *whole* of the raw data, when the prior element is

$$d\mu\, d\Sigma/|\Sigma|^{\frac{1}{2}(p+1)}$$

(Geisser, 1965). Hence the posterior distribution of Φ_1 is obtained by standardizing Σ_1, the corresponding sub-matrix of Σ, where Σ has distribution $IW(\mathbf{S}; \nu, p)$. By the theory of the Wishart distribution, Σ_1 has distribution $IW(\mathbf{S}_1; \nu-(p-q), p)$ where $\mathbf{S}_1$ is the appropriate sub-matrix of $\mathbf{S}$. Thus, if Ψ_1 is the standardized inverse of Φ_1, the posterior density of Ψ_1 is given by Fisher's formula to be

$$\pi(\Psi_1|\mathbf{R}) = \pi(\Psi_1|\mathbf{S}) = \propto |\mathbf{R}_1|^{\frac{1}{2}\nu_1}|\Psi_1|^{\frac{1}{2}(\nu_1-q-1)}F_{\nu_1}(\Gamma_1), \tag{3.5}$$

where $\nu_1 = \nu-p+q$, $\mathbf{R}_1$ is the appropriate sub-matrix of $\mathbf{R}$, and Γ_1 is $(q \times q)$, with $(\Gamma_1)_{ij} = (\Psi_1)_{ij}r_{ij}$.

So the posterior distribution of Φ_1 involves only $\mathbf{R}_1$, while the kernel of the likelihood based on $\mathbf{R}_1$ is $|\Psi_1|^{\frac{1}{2}\nu}F_\nu(\Gamma_1)$, which is not a factor of (3.5). Thus the prior $d\Phi/|\Phi|^{\frac{1}{2}(p+1)}$, which appears to be the only one for which the posterior distribution of Φ_1 involves $\mathbf{R}_1$ alone, inevitably leads to a marginalization paradox for Φ_1.

3.2. *Groups in the Background—"A Grin Without a Cat!"*

We now investigate more deeply the structure of the paradox in Examples 5 and 6.

We have observable data x dependent on a parameter θ, and functions z, ζ of x, θ respectively such that the distribution of z depends only on ζ. Section 2 presented conditions involving a group structure for the problem under which we can find a prior distribution for θ such that the posterior distribution for ζ depends only on z. Examples 5 and 6 show that these conditions are not necessary. The conditions we investigate below are presumably not necessary either but are sufficient to cover Examples 5 and 6 and much else besides.

We suppose that the observable data x is itself a function of (perhaps fictitious) raw data $\hat{x} \in \hat{X}$. The distribution of $\hat{x}$ depends on a parameter $\hat{\theta} \in \hat{\Theta}$, and θ is a function of $\hat{\theta}$. Within this extended structure, we suppose that the problem of inference about θ is constant under exact transformation groups G on $\hat{X}$ and $\bar{G}$ on $\hat{\Theta}$, and that x is a maximal invariant under the action of G on $\hat{X}$. It then follows from Lemma 2.1 that the distribution of x does depend only on θ. We further suppose that the problem of inference about ζ is constant under exact transformation groups T on $\hat{X}$ and $\bar{T}$ on $\hat{\Theta}$, with z a maximal invariant under T, so that we have the distribution of z depending on ζ alone.

We can represent $\hat{X} = Y \times Z$, $\hat{\Theta} = \mathcal{Y} \times \mathcal{Z}$ in the usual way, where $Y = T$, $\mathcal{Y} = \bar{T}$, $Z \subset \hat{X}$, $\mathcal{Z} \subset \hat{\Theta}$ and z, ζ may be regarded as taking values in $Z, \mathcal{Z}$ respectively. It is clear that G is a subgroup of T, so that G acts on Y as an exact transformation group, with $g \circ y = gy$. Hence we can further represent $Y = \hat{Y} \times W$ with $\hat{Y} = G$ and $W \subset T$. Then $\hat{X} = \hat{Y} \times W \times Z$ where we may identify $W \times Z$ with X. Similarly,

$$\hat{\Theta} = \hat{\mathcal{Y}} \times \Omega \times \mathcal{Z} = \hat{\mathcal{Y}} \times \Theta$$

with $\hat{\mathcal{Y}} = \bar{G}, \Omega \subset \bar{T}, \mathcal{Z} \subset \Theta$.

We shall pass backwards and forwards between the equivalent representations

$$\hat{x} = (y, z) = (\hat{y}, w, z) = (\hat{y}, x),$$
$$\hat{\theta} = (\eta, \zeta) = (\hat{\eta}, \omega, \zeta) = (\hat{\eta}, \theta).$$

By Lemma 2.1, we have density elements of the form

$$P_{\hat{\theta}}(d\hat{x}) = f_0(y \mid z;\ \hat{\theta})\, f_3(z \mid \zeta)\, \mu_T(dy)\, dz.$$

Also, because of the constancy (Section 2.1) under G and $\bar{G}$ it is easy to see that

$$f_0(y \mid z;\ \hat{\theta})\, \mu_T(dy) = f_1(\hat{y} \mid x;\ \theta) f_2(w \mid z;\ \theta)\, \mu_G(d\hat{y})\, dw. \tag{3.6}$$

Hence

$$P_{\hat{\theta}}(d\hat{x}) = f_1(\hat{y} \mid x;\ \hat{\theta})\, \mu_G(d\hat{y}).f_2(w \mid z;\ \theta)\, dw.f_3(z \mid \zeta)\, dz \tag{3.7}$$

and so the density element for the observed data $x = (w, z)$ is

$$f_2(w \mid z;\ \theta)\, dw.f_3(z \mid \zeta)\, dz. \tag{3.8}$$

We now attempt to find a prior distribution of θ for which the posterior for ζ depends only on z, within the large class of distributions with density element of the form $\pi(\omega, \zeta)\, d\omega\, d\zeta$. Here $d\zeta$ is an arbitrary measure element, while the element $d\omega$ is inherited from the grafted group structure by noticing that, since $\mu_{\bar{T}}$ is a left-invariant measure under the operation of $\bar{G}$ on $\mathcal{Y}$, we must have a product representation

$$\mu_{\bar{T}}(d\eta) = \mu_{\bar{G}}(d\hat{\eta})\, d\omega. \tag{3.9}$$

With the above prior, we find the marginal posterior distribution for ζ to have element proportional to

$$d\zeta . \int_\Omega f_2(w \mid z;\ \omega, \zeta) f_3(z \mid \zeta)\, \pi(\omega, \zeta)\, d\omega. \tag{3.10}$$

The constancy under G and $\bar{G}$ implies that, for any $g \in G$,

$$1 = \int_G f_1(\hat{y} \mid x;\ \hat{\eta}, \theta)\, \mu_G(d\hat{y}) = \int_G f_1(g \mid x;\ \bar{g}\bar{\mathcal{Y}}^{-1}\hat{\eta}, \theta)\, \mu_G(d\hat{y})$$

$$= k\Delta_{\bar{G}}(\bar{g}) \int_{\bar{G}} f_1(g \mid x;\ \hat{\eta}', \theta)\, \nu_{\bar{G}}(d\hat{\eta}') \quad \text{for some constant } k.$$

Thus (3.10) is proportional to

$$d\zeta . \int_\Omega \int_{\bar{G}} f_1(g \mid x;\ \hat{\eta}, \omega, \zeta) f_2(w \mid z;\ \omega, \zeta) f_3(z \mid \zeta) .\ \pi(\omega, \zeta)\, \Delta_{\bar{G}}(\bar{g}) \{\Delta_{\bar{G}}(\hat{\eta})\}^{-1}\, d\omega\, \mu_{\bar{G}}(d\hat{\eta})$$

$$\propto f_3(z \mid \zeta)\, d\zeta . \int_{\bar{T}} f_0(t \mid z;\ \eta, \zeta)\, \delta(\eta, \zeta)\, \mu_{\bar{T}}(d\eta) .\ \Delta_{\bar{G}}(\bar{g}), \tag{3.11}$$

where $g \in G$ is arbitrary, $t = gw$, $\eta = \hat{\eta}\omega$ and $\delta(\eta, \zeta) = \{\Delta_G(\hat{\eta})\}^{-1}\pi(\omega, \zeta)$; and this may finally be reduced to

$$f_3(z|\zeta)\,d\zeta \cdot \int_{\bar{T}} f_0(e|z;\eta,\zeta)\,\delta(\tilde{\iota}\eta,\zeta)\,\mu_{\bar{T}}(d\eta) \cdot \Delta_{\bar{G}}(\bar{g}) \tag{3.12}$$

since the extended problem is constant under T and $\bar{T}$.

If (3.10) is to yield a posterior distribution which does not depend on w, it must be of the form

$$p(w,z)\,q(z,\zeta)\,d\zeta \tag{3.13}$$

and so using (3.12), for all $t \in T$,

$$\int_{\bar{T}} f_0(e|z;\eta,\zeta)\,\delta(\tilde{\iota}\eta,\zeta)\,\mu_{\bar{T}}(d\eta)$$

must be of the form

$$p(t,z)\,q'(z,\zeta) \tag{3.14}$$

where

$$p(t,z) = \{\Delta_{\bar{G}}(\bar{g})\}^{-1}p(w,z).$$

One case in which (3.14) will hold is when $\delta(\eta, \zeta)$ is of the form $a(\zeta)\beta(\eta)$, where β is a morphism on $\bar{T}$. For then the integral becomes

$$\beta(\tilde{\iota})\,a(\zeta)\int_{\bar{T}} f_0(e|z;\eta,\zeta)\,\beta(\eta)\,\mu_{\bar{T}}(d\eta)$$

which is of the desired form. It is plausible that in problems with special structure (3.14) may hold even when $\delta(\eta, \zeta)$ is not of this form, but it seems a reasonable conjecture that this condition will normally be necessary, and so we proceed on this assumption.

If the prior distribution is to yield the relation $\delta(\eta, \zeta) \equiv a(\zeta)\beta(\eta)$ we find

$$\pi(\omega,\zeta) \equiv \Delta_{\bar{G}}(\hat{\eta})\,\beta(\hat{\eta})\,\beta(\omega)\,a(\zeta)$$

and it follows that $\beta(\hat{\eta}) \equiv \{\Delta_{\bar{G}}(\hat{\eta})\}^{-1}$, while $\pi(\omega, \zeta) = a(\zeta)\beta(\omega)$.

Thus, in order to have a posterior marginal distribution for ζ dependent only on z, for a prior in the family investigated, it will usually be necessary to use a prior density element $d\theta \propto \beta(\omega)\,d\omega\,d\zeta$, where $d\zeta$ has been re-defined but is still arbitrary; but β *must* be an extension of the morphism $\{\Delta_{\bar{G}}(\bar{g})\}^{-1}$ from $\bar{G}$ to $\bar{T}$.

It remains to consider whether such a prior may be used without paradoxical effect. The following argument shows that this is frequently impossible.

Note that

$$\beta(\eta)\,\mu_{\bar{T}}(d\eta)\,d\zeta = \beta(\hat{\eta})\,\mu_{\bar{G}}(d\hat{\eta})\,\beta(\omega)\,d\omega\,d\zeta = \{\Delta_{\bar{G}}(\hat{\eta})\}^{-1}\mu_{\bar{G}}(d\hat{\eta})\,d\theta \tag{3.15}$$

and, as the analysis of Section 2.4 demonstrates, this is the prior over $\hat{\mathcal{Y}} \times \Theta$ for which no paradox arises with regard to θ; that is, we get the *same* posterior distribution for θ whether we use the prior (3.15) and then marginalize, or whether we use the prior element $\beta(\omega)\,d\omega\,d\zeta$ and the sampling density of x given θ.

It follows that the marginal posterior of ζ in this restricted case must be the same as we would get by using the full group-structure model and the prior (3.15). But, as Section 2.4 and Appendix 1 demonstrate, the choice of this prior distribution will

often entail a marginalization paradox for ζ, as in Examples 5 and 6. Then, although it is possible to use some prior $d\theta$ giving posterior inferences for ζ depending only on z, it will not be possible to do so in a non-paradoxical way, at least within the class of priors investigated. As Appendix 2 demonstrates, this class may be extendible to the class of all non-trivial prior distributions.

4. MARGINALIZATION IN STRUCTURAL INFERENCE

In this section we renew our contact with Fraser's theory of structural inference, a concise summary of the relevant portions of which is to be found in Appendix 3.

We will examine a modification of the "progression model" of Example 3. In deference to Fraser's theory, our arguments will be stated entirely within the ambit of that theory and the strong connections that exist between structural inference and Bayesian theory (Appendix 3) will here be ignored. In this way, we will demonstrate that the theory of structural inference is powerful enough to develop its own paradoxes without the assistance of improper Bayesians.

Example 7. An illustration of the structure of Example 3 is provided by a bivariate-normal generator represented by the model

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \beta & \sigma_2 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{4.1}$$

in which e_1 and e_2 are independent, internally generated $N(0,1)$ error variables and $\beta, \sigma_1, \sigma_2$ are fixed but unknown parameters $(-\infty<\beta<\infty, \sigma_1>0, \sigma_2>0)$. We have chosen the parametrization $\theta=(\beta,\sigma_1,\sigma_2)$, rather than that of Example 3, because (4.1) is then suitable for analysis by a statistician, F, who is required to make structural inference about θ on the basis of n independent replications of (4.1),

$$x = \left\{ \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}, \ldots, \begin{pmatrix} x_{n1} \\ x_{n2} \end{pmatrix} \right\}.$$

Such inference is possible because the transformation matrices in (4.1) constitute a group.

Application of the "structural" theory to (4.1) would give a structural distribution for θ having density element

$$\pi(d\beta, d\sigma_1, d\sigma_2 \mid x) \propto (\sigma_1\sigma_2)^{-n} \exp\left\{ -\frac{1}{2}\frac{\sum x_{i1}^2}{\sigma_1^2} - \frac{1}{2}\frac{\sum (x_{i2}-\beta\sigma_1^{-1}x_{i1})^2}{\sigma_2^2} \right\} d\beta \frac{d\sigma_1}{\sigma_1^2}\frac{d\sigma_2}{\sigma_2}. \tag{4.2}$$

(As asserted in Section 1, Example 3, this agrees with (1.8).) Before F is able to make this inference, he is given some additional information about the generator (Fig. 1).

Two independently rotating pointers are operated separately by two technicians T_1 and T_2 to generate two random deviates, u_1 and u_2, independently and uniformly distributed on $(0,1)$. These deviates are fed into the Box–Muller transformer (1958) to yield e_1, e_2 by the fixed transformation

$$\left.\begin{aligned} e_1 &= (-2\log_e u_1)^{\frac{1}{2}} \cos 2\pi u_2, \\ e_2 &= (-2\log_e u_1)^{\frac{1}{2}} \sin 2\pi u_2. \end{aligned}\right\} \tag{4.3}$$

T_1 and T_2 are instructed to set the pointers initially to zero on the circumferential scale, which is marked from 0 to 1. Each then has to give his pointer a vigorous twist

to generate the first pair (u_{11}, u_{12}), which is then passed through the Box–Muller transformer (4.3) and triangular transformation (4.1) to give (x_{11}, x_{12}) for F's inspection. Successive twists are made with the pointers starting at the immediately previous resting positions, so that the uniform deviates

$$\left\{\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}, \ldots, \begin{pmatrix} u_{n1} \\ u_{n2} \end{pmatrix}\right\}$$

are sequentially generated and transformed into x. This more detailed description of the generator does not in itself change F's satisfaction with (4.2).

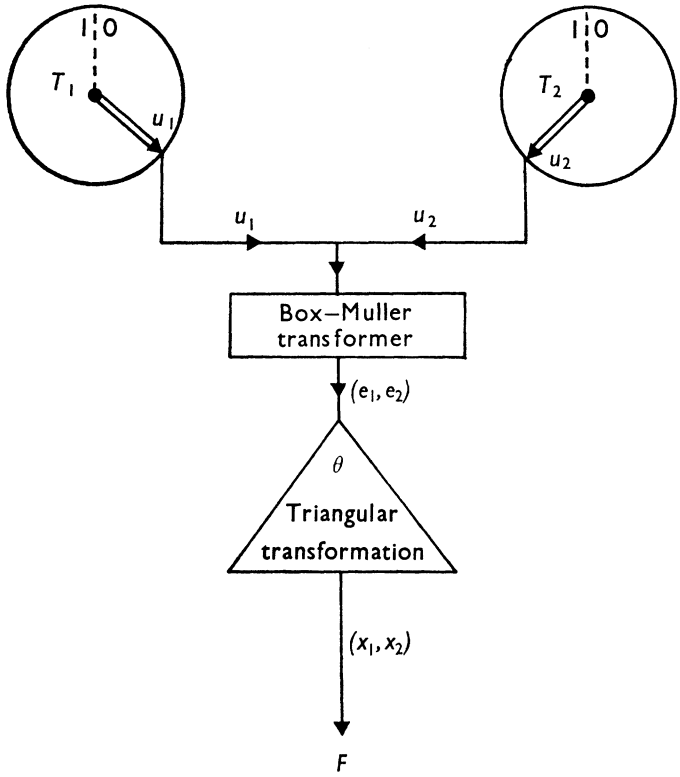

FIG. 1. The bivariate normal generator.

However, T_2 then informs F that, for the data under analysis, he neglected to set his pointer to zero for the first twist. Suppose that the unknown scale reading of the initial position of this pointer is $\lambda/2\pi$. Under the "classical model of statistics" (Fraser, 1968, p. 185), it is clear that this information of T_2's could be ignored, since the distribution of x conditional on any θ is independent of λ. However, F is obliged to incorporate the information into his analysis since the unknown initial starting angle λ is not part of the basic physical internal error but rather represents an unknown transformation of it. Thus we write

$$u_2 = \lambda/2\pi + u_2' \quad \text{(modulo 1)}, \tag{4.4}$$

where u_2' is the physical deviate determined by the movement of the pointer from its initial position on each of the n successive twists. With (4.4), the Box–Muller transformer gives

$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} \cos\lambda & -\sin\lambda \\ \sin\lambda & \cos\lambda \end{pmatrix} \begin{pmatrix} e_1' \\ e_2' \end{pmatrix}, \tag{4.5}$$

where

$$e_1' = (-2\log_e u_1)^{\frac{1}{2}} \cos 2\pi u_2',$$

$$e_2' = (-2\log_e u_1)^{\frac{1}{2}} \sin 2\pi u_2'.$$

So our extended model incorporating T_2's information is

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \beta & \sigma_2 \end{pmatrix} \begin{pmatrix} \cos\lambda & -\sin\lambda \\ \sin\lambda & \cos\lambda \end{pmatrix} \begin{pmatrix} e_1' \\ e_2' \end{pmatrix} \tag{4.6}$$

in which e_1', e_2' are independent $N(0,1)$ variables, generated from the basic physical rotations of the pointers. With λ completely unknown, (4.6) is actually a structural model, $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ being obtained from $\begin{pmatrix} e_1' \\ e_2' \end{pmatrix}$ by a general linear transformation. F finds that the structural distribution for the extended parameter $(\lambda, \beta, \sigma_1, \sigma_2)$ has density element

$$\pi(d\lambda, d\beta, d\sigma_1, d\sigma_2) \propto (\sigma_1 \sigma_2)^{-n} \exp\left\{ -\frac{1}{2}\frac{\sum x_{i1}^2}{\sigma_1^2} - \frac{1}{2}\frac{\sum (x_{i2} - \beta\sigma_1^{-1}x_{i1})^2}{\sigma_2^2} \right\} d\lambda\, d\beta \frac{d\sigma_1}{\sigma_1}\frac{d\sigma_2}{\sigma_2^2}. \tag{4.7}$$

The interpretation of (4.7) in structural terms is straightforward. The conditional distribution of λ given $(\beta, \sigma_1, \sigma_2)$ is uniform on $(0, 2\pi)$; this conditional distribution is both the conditional distribution as usually defined from a joint distribution in probability theory and also the structural distribution that is derivable from

$$\begin{pmatrix} \sigma_1 & 0 \\ \beta & \sigma_2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \cos\lambda & -\sin\lambda \\ \sin\lambda & \cos\lambda \end{pmatrix} \begin{pmatrix} e_1' \\ e_2' \end{pmatrix},$$

available when $(\beta, \sigma_1, \sigma_2)$ is assumed known. The conditional uniformity of λ is surely acceptable. However, the marginal distribution of $(\beta, \sigma_1, \sigma_2)$ determined by (4.7) differs from that given by (4.2). T_2's information has after all effected a change in F's inference!

T_1 now confirms T_2's information with the remark that he had observed that the initial setting of T_2's needle had been *different* from zero, although he could not now give any idea about the particular value it took. The exclusion of the value $\lambda = 0$ resulting from this information has the immediate consequence that the set of transformations in (4.6) is no longer a group.

If F were to continue to follow the methods of Fraser, he would be obliged to employ *conditional analysis* (Fraser, 1968, Chapter 4). This is achieved by re-writing (4.6) in the form (4.1) with

$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} \cos\lambda & -\sin\lambda \\ \sin\lambda & \cos\lambda \end{pmatrix} \begin{pmatrix} e_1' \\ e_2' \end{pmatrix},$$

that is, the distribution of the error in (4.1) now depends, formally at least, on the parameter λ, the *additional quantity* of Fraser (1968, p. 188). Now it happens that,

conditional on λ, the structural distribution of $(\beta, \sigma_1, \sigma_2)$ is given by (4.2), the distribution F was proposing to adopt before the technicians interrupted him. T_1's information, which was of an apparently trivial nature, has had a major effect on the structural inference.

5. DISCUSSION

Why should we consider the marginalization paradoxes of Section 1 important? It cannot be claimed that the full implications of the paradox have been explored and appreciated. One implication that can be made reasonably explicit concerns a generalization of a theorem of Dawid and Stone (1972): unless the posterior density element $\pi(d\zeta|z)$ has a Bayesian kernel $\pi(d\zeta)f(z|\zeta)$, say, the generalized theorem will show $\pi(d\zeta|z)$ to be *expectation inconsistent* with $f(z|\zeta)$. Roughly speaking, this means that a system of bets on (z, ζ) can be found that is fair when evaluated by $\pi(d\zeta|z)$ for each z but is uniformly unfair when evaluated by $f(z|\zeta)$ for each ζ. A closely related result in decision theory is that, in decision problems in which the loss is determined by ζ, for decision functions based on z to be admissible, they must be derivable from a Bayesian kernel of the above type (Sacks, 1963).

The main result of Section 2.3 may be interpreted as an argument for right-invariant prior measures. General support for such measures is provided by structural theory (Fraser, 1961), the theory of approximability of invariant posterior distributions by proper priors (Stone, 1970) and the requirements for best equivariant procedures in decision theory (Zidek, 1969). Much current Bayesian practice implicitly, if not explicitly, employs right-invariant measures to represent ignorance (Lindley, 1965). Moreover, invariant Bayesian confidence regions possess the classical confidence region property, if constructed using right-invariant priors (Stein, 1965; Hora and Buehler, 1966). However, such general support does not resolve the difficulty, revealed in Section 2.4, where two "rights" appear to make a "wrong"; that is, the choice of right invariant prior on $\bar{T}$ may well induce the paradox which right invariant prior on $\bar{G}$ avoids.

In fact the resolution of this conflict *can* be made in terms of the approximability theory already cited. The form of the resolution is exemplified in Stone and Dawid (1972); it is asymptotically attainable only as the data considered approach the limiting type of data that is consistent with use of right-invariant prior.

We have not succeeded in finding an index to measure the degree of inconsistency when the paradox is present. It is obvious that, in cases involving a large number n of replicate observations, the inconsistency will be relatively unimportant, by the "principle of precise measurement" (Edwards *et al.*, 1963).

The examples of Sections 1 and 3 relate directly to a general statistical problem that has received previous attention (Fisher, 1956; Cox, 1958; Kalbfleisch and Sprott, 1970). With appropriate notation, the examples provide illustrations of the following decomposition:

$$f(x|\theta) = f(z|\zeta)f(y|z;\ \zeta, \xi), \tag{5.1}$$

where $x = (y, z)$ and $\theta = (\zeta, \xi)$. In the cases of (5.1) investigated by Kalbfleisch and Sprott (1970), there is alleged to be "no available information concerning ζ in the second factor on the right-hand side of (5.1), in the absence of knowledge about ξ" and inference is based on $f(z|\zeta)$. As indicated by Smith (1970), it is doubtful whether precise and consistent interpretation of the quotation is possible. However, the motivation for restriction of the data to z is clear except to the most committed

Bayesian (such as B_1!). With such restriction, it is necessary to specify a prior distribution only for ζ (which appeals to B_2!). From B_1's viewpoint, problems having the structure (5.1) may be divided into two categories: *reducible*, for which there exist prior measures for θ such that the posterior distribution of ζ is determined by z only, and *irreducible*, for which no such prior exists. The first category has been illustrated by each of Examples 1–6.

For a problem in the reducible category, the prior measures for θ that do result in dependence on z only, as far as inference about ζ is concerned, may be divided into two classes: *paradoxical* (following the pattern established) and *paradox-free* (where B_1 can agree with B_2). A paradoxical prior is surely a bad choice and, presumably (as in Examples 1–3) B_1 will be led to choose a paradox-free prior for θ. Examples 4a and 4b show that there may be no such choice that suffices for two factorizations of the form (5.1) of a given $f(x \mid \theta)$. Even worse, Example 5 shows that the paradox-free class may be empty; we confidently conjecture the same for Example 6.

We note in passing that Example 4a is relevant to the problem of the ratio of normal means, which is equivalent to ζ_1/ζ_2. The paradoxical prior measure $d\zeta_1 d\zeta_2$ is just that required to generate the well-known Creasy solution (Creasy, 1954).

Choice of a paradoxical prior is definitely "un-Bayesian". For instance, in Example 6, inference about a principal sub-matrix of a correlation matrix based on only the corresponding sample sub-matrix must be unBayesian; roughly speaking, the "correlations" between the sample correlation coefficients should not be neglected. However, description of paradox-free priors as "Bayesian" is premature as Examples 4a and 4b demonstrate.

A technical problem, suggested by Examples 5 and 6, is whether, for the decomposition (5.1), there always (or nearly always) exists a group-structural extension.

The difficulties revealed for Fraser's structural theory in Section 4 are related to those for the Bayesian analyses because they share a common mathematical root, the fact that right-invariant measure on a group need not induce right-invariant measure on a sub-group (see Section 2.4). These paradoxes appear to be an inevitable price paid by a theory that becomes too closely attached to a mathematical structure, here a group, without extensive investigation of the statistical consequences of such attachment.

The particular problem considered in Section 4 involves specializations of more general models that occupy a central position in Fraser's book, namely, the "progression model" in Chapter 3 and the multivariate model in Chapter 5. Further work has been reported by Fraser and Haq (1969) while a related paper is Villegas (1971). A specific point brought out in Section 4 is that Fraser's theory is inconsistent in the following sense. If the group structure for λ in (4.6) when $0 \leqslant \lambda < 2\pi$ is ignored and the conditional analysis, used when $0 < \lambda < 2\pi$, is applied, the result is inconsistent with what the structural theory gives when the group structure is taken into account. This suggests that the adjacent layers of the theory are in basic conflict.

REFERENCES

BONDAR, J. V. (1972). Structural distributions without exact transitivity. *Ann. Math. Statist.*, **43**, 326–339.

BOX, G. E. P. and MULLER, M. E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.*, **29**, 610–611.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, **29**, 357–372.
CREASY, M. A. (1954). Limits for the ratio of means. *J. R. Statist. Soc. B*, **16**, 186–194.
DAWID, A. P. and STONE, M. (1972). Expectation consistency of inverse probability distributions. *Biometrika*, **59**, 486–489.
DEMPSTER, A. P. (1969). *Elements of Continuous Multivariate Analysis*. Reading, Mass.: Addison-Wesley.
EDWARDS, W., LINDMAN, H. and SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.
FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. London: Oliver & Boyd.
—— (1962). The simultaneous distribution of correlation coefficients. *Sankhyā*, **24**, 1–8.
FRASER, D. A. S. (1961). On fiducial inference. *Ann. Math. Statist.*, **32**, 661–676.
—— (1968). *The Structure of Inference*. New York: Wiley.
FRASER, D. A. S. and HAQ, M. S. (1969). Structural probability and prediction for the multivariate model. *J. R. Statist. Soc. B*, **31**, 317–331.
GEISSER, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.*, **36**, 150–159.
GEISSER, S. and CORNFIELD, J. (1963). Posterior distributions for multivariate normal parameters. *J. R. Statist. Soc.* B, **25**, 368–376.
HARTIGAN, J. (1964). Invariant prior distributions. *Ann. Math. Statist.*, **35**, 836–845.
HORA, R. B. and BUEHLER, R. J. (1966). Fiducial theory and invariant estimation. *Ann. Math. Statist.*, **37**, 643–656.
JEFFREYS, H. (1961). *Theory of Probability*, 3rd edn. Oxford: Clarendon.
KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with Discussion). *J. R. Statist. Soc.* B, **32**, 175–208.
LINDLEY, D. V. (1965). *Introduction to Probability and Statistics. Part* 2: *Inference*. Cambridge: University Press.
LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *J. R. Statist. Soc.* B, **34**, 1–41.
NACHBIN, L. (1965). *The Haar Integral*. New York: Van Nostrand.
NOVICK, M. R. (1969). Multiparameter Bayesian indifference procedures. *J. R. Statist. Soc.* B, **31**, 29–51.
RAIFFA, H. A. and SCHLAIFER, R. S. (1961). *Applied Statistical Decision Theory*. Boston: Graduate School of Business Administration, Harvard University.
SACKS, J. (1963). Generalized Bayes solutions in estimation problems. *Ann. Math. Statist.*, **34**, 751–768.
SMITH, A. F. M. (1970). In discussion of Kalbfleisch, J. D. and Sprott, D. A. (1970).
STEIN, C. M. (1965). Approximation of improper prior measures by prior probability measures. In *Bernoulli, Bayes, Laplace* (J. Neyman and L. Le Cam, eds), pp. 217–240. Berlin: Springer.
STONE, M. (1970). Necessary and sufficient condition for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.*, **41**, 1349–1353.
STONE, M. and DAWID, A. P. (1972). UnBayesian implications of improper Bayes inference in routine statistical problems. *Biometrika*, **59**, 369–375.
VILLEGAS, C. (1971). On Haar priors. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.), pp. 409–414. Toronto, Montreal: Holt, Rinehart and Winston.
WILKINSON, G. N. (1971). In discussion of Godambe, V. P. and Thompson, M. E. (1971). Bayes, fiducial and frequency aspects of statistical inference in regression analysis in survey-sampling. *J. R. Statist. Soc.* B, **33**, 361–376.
ZELLNER, A. (1971). *An Introduction to Bayesian Statistics in Econometrics*. New York: Wiley.
ZIDEK, J. V. (1969). A representation of Bayes invariant procedures in terms of Haar measure. *Ann. Inst. Statist. Math.*, **21**, 291–308.

APPENDIX 1

Estimation with a Wishart Distribution

The multivariate normal distribution has a special structure which leads to an absence of marginalization paradoxes in cases where we might expect them. This Appendix tries to pinpoint those aspects of the structure which are responsible, while showing that the paradox persists for some types of inference.

Let $(\mathbf{x}_1, \ldots, \mathbf{x}_\nu)$ $(\nu > p+q)$ be a random sample from the $(p+q)$-dimensional normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is non-singular but unknown. The matrix $\mathbf{S} = \sum_{i=1}^{\nu} \mathbf{x}_i \mathbf{x}_i'$ is sufficient for $\boldsymbol{\Sigma}$, and has the Wishart distribution $W(\boldsymbol{\Sigma}, \nu, p+q)$.

The family of distributions for $\mathbf{x}$ is invariant under the *general linear group* T of all non-singular $(p+q)\times(p+q)$ matrices. That is, if $\mathbf{M} \in T$, and we put $\mathbf{M} \circ \mathbf{x} = \mathbf{M}\mathbf{x}$, $\mathbf{M} \circ \mathbf{S} = \mathbf{M}\mathbf{S}\mathbf{M}'$, $\mathbf{M} \circ \boldsymbol{\Sigma} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}'$, then $\mathbf{M} \circ \mathbf{x}_i \sim N(\mathbf{0}, \mathbf{M} \circ \boldsymbol{\Sigma})$, or equivalently

$$\mathbf{M} \circ \mathbf{S} \sim W(\mathbf{M} \circ \boldsymbol{\Sigma}, \nu, p+q).$$

We shall be particularly interested in the class of prior distributions for $\boldsymbol{\Sigma}$ which are relatively invariant under T, since this includes most of the recommended "ignorance priors". Such distributions have density element proportional to

$$d\boldsymbol{\Sigma} / |\boldsymbol{\Sigma}|^{\frac{1}{2}\nu} \tag{A1.1}$$

for some value of ν.

Let $\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix}$, where $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ have respectively p and q components, and correspondingly partition

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Define $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, the residual covariance matrix and the matrix of regression coefficients of the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$. Similarly, define $\mathbf{S}_{11.2}$ and $\mathbf{D}$ from $\mathbf{S}$. Let $\boldsymbol{\Phi}_{22}, \boldsymbol{\Phi}_{11.2}, \mathbf{R}_{22}, \mathbf{R}_{11.2}$ be the correlation matrices obtained by standardizing $\boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_{11.2}, \mathbf{S}_{22}, \mathbf{S}_{11.2}$ respectively. We shall be interested in inferences about $\boldsymbol{\Sigma}_{22}, \boldsymbol{\Phi}_{22}, \boldsymbol{\Sigma}_{11.2}, \boldsymbol{\Phi}_{11.2}$.

It may be verified that (A1.1) is equivalent to a density element

$$(d\boldsymbol{\Sigma}_{22} / |\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}(\nu-2p)}) \cdot (d\boldsymbol{\Sigma}_{11.2} / |\boldsymbol{\Sigma}_{11.2}|^{\frac{1}{2}\nu}) \, d\boldsymbol{\Delta}. \tag{A1.2}$$

The special nature of the Wishart distribution is embodied in the factorization of the density:

$$\begin{aligned} &f(\mathbf{S}_{22}, \mathbf{S}_{11.2}, \mathbf{D} | \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_{11.2}, \boldsymbol{\Delta}) \, d\mathbf{S}_{22} \, d\mathbf{S}_{11.2} \, d\mathbf{D} \\ &\quad = f(\mathbf{S}_{22} | \boldsymbol{\Sigma}_{22}) \, d\mathbf{S}_{22} \cdot f(\mathbf{S}_{11.2} | \boldsymbol{\Sigma}_{11.2}) \, d\mathbf{S}_{11.2} \cdot f(\mathbf{D} | \mathbf{S}_{22}; \boldsymbol{\Sigma}_{11.2}, \boldsymbol{\Delta}) \, d\mathbf{D} \end{aligned} \tag{A1.3}$$

(Dempster, 1969, p. 296).

(i) *Inference about* $\boldsymbol{\Sigma}_{22}$. The problem of inference about $\boldsymbol{\Sigma}_{22}$ remains constant under the group G_1 of transformations of the form

$$\mathbf{S} \to \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \circ \mathbf{S}, \quad \boldsymbol{\Sigma} \to \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \circ \boldsymbol{\Sigma},$$

where $\mathbf{A}$ is upper triangular with positive diagonal elements. This group is exact, and induces the decompositions

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11.2}^{\frac{1}{2}} & \mathbf{D} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \circ \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11.2}^{\frac{1}{2}} & \boldsymbol{\Delta} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \circ \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{M}^{\frac{1}{2}}$ denotes the (unique) upper triangular matrix with positive diagonal elements such that $\mathbf{M}^{\frac{1}{2}}(\mathbf{M}^{\frac{1}{2}})' = \mathbf{M}$.

By the theory of Section 2, the distribution of $\mathbf{S}_{22}$ depends only on $\boldsymbol{\Sigma}_{22}$ (as (A1.3) states) and any prior distribution relatively invariant under G_1 will yield a marginal posterior for $\boldsymbol{\Sigma}_{22}$ involving only $\mathbf{S}_{22}$. However, it is not necessary to take a prior for $\boldsymbol{\Sigma}$ which induces right-invariant measure on $\begin{pmatrix} \boldsymbol{\Sigma}_{11.2}^{\frac{1}{2}} & \boldsymbol{\Delta} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, in order to avoid a marginalization paradox. It is easily seen from (A1.3) that any prior of the form

$$\pi(\boldsymbol{\Sigma}_{22})\,\pi(\boldsymbol{\Sigma}_{11.2}, \boldsymbol{\Delta})\,d\boldsymbol{\Sigma}_{22}\,d\boldsymbol{\Sigma}_{11.2}\,d\boldsymbol{\Delta} \tag{A1.4}$$

will suffice. In particular, any prior (A1.1) will avoid a paradox for $\boldsymbol{\Sigma}_{22}$, since (A1.2) has this form.

(ii) *Inference about* $\boldsymbol{\Phi}_{22}$. Since $\boldsymbol{\Phi}_{22}$ is a function of $\boldsymbol{\Sigma}_{22}$, its posterior distribution may be found from that of $\boldsymbol{\Sigma}_{22}$. For a prior of the form (A1.4), this posterior distribution for $\boldsymbol{\Sigma}_{22}$ is just that obtained using the likelihood based on $\mathbf{S}_{22}$, which is $W(\boldsymbol{\Sigma}_{22}, \nu, q)$, with the prior element $\pi(\boldsymbol{\Sigma}_{22})\,d\boldsymbol{\Sigma}_{22}$. In this smaller problem, inference about $\boldsymbol{\Phi}_{22}$ is constant under the group $\mathbf{S}_{22} \to \mathbf{L} \circ \mathbf{S}_{22}$, i.e. $\mathbf{L}\mathbf{S}_{22}\mathbf{L}', \boldsymbol{\Sigma}_{22} \to \mathbf{L} \circ \boldsymbol{\Sigma}_{22}$, where $\mathbf{L} = \operatorname{diag}(l_1, \ldots, l_q)$ with $l_i > 0$. We have corresponding decompositions $\mathbf{S}_{22} = \mathbf{U}\mathbf{R}_{22}\mathbf{U}'$, $\boldsymbol{\Sigma}_{22} = \mathbf{Y}\boldsymbol{\Phi}_{22}\mathbf{Y}'$ where

$$\mathbf{U} = \operatorname{diag}(s_{p+1}, \ldots, s_{p+q}),$$

$$\mathbf{Y} = \operatorname{diag}(\sigma_{p+1}, \ldots, \sigma_{p+q}),$$

where $s_i = s_{ii}^{\frac{1}{2}}, \sigma_i = \sigma_{ii}^{\frac{1}{2}}$. Relatively invariant measures under this group will be of the form

$$\pi(\boldsymbol{\Sigma}_{22})\,d\boldsymbol{\Sigma}_{22} = \prod_{i=p+1}^{p+q} \frac{d\sigma_i}{\sigma_i^{a_i}}\,\pi(\boldsymbol{\Phi}_{22})\,d\boldsymbol{\Phi}_{22}$$

where a_i is arbitrary. In this case we do require right-invariant measure on T to avoid a paradox. This has $a_i \equiv 1$, and is also left-invariant. However, the element

$$d\boldsymbol{\Sigma}_{22} / |\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}(v-2p)},$$

implied by (A1.1), is relatively invariant with

$$d(\mathbf{M} \circ \boldsymbol{\Sigma}_{22}) / |\mathbf{M} \circ \boldsymbol{\Sigma}_{22}|^{\frac{1}{2}(v-2p)} = |\mathbf{M}|^{q+2p-v+1}\,d\boldsymbol{\Sigma}_{22} / |\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}(v-2p)}$$

so that it is left-invariant under transformations of the form $\mathbf{L}$ only when

$$|\mathbf{L}|^{q+2p-v+1} = 1$$

for all $\mathbf{L}$, and this implies $v = q+2p+1$. It follows that a marginalization paradox will arise for $\boldsymbol{\Phi}_{22}$, with prior (A1.1), unless $v = p+(p+q+1)$. Clearly the paradox cannot be simultaneously avoided for two different values of p.

(iii) *Inference about* $\boldsymbol{\Sigma}_{11.2}$. The problem of inference about $(\boldsymbol{\Sigma}_{11.2}, \boldsymbol{\Sigma}_{22})$ together is constant under the group G_2 of transformations of the form $\begin{pmatrix} \mathbf{I} & \mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, which induces

the decompositions

$$\mathbf{S} = \begin{pmatrix} \mathbf{I} & \mathbf{D} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \circ \begin{pmatrix} \mathbf{S}_{11.2} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{I} & \mathbf{\Delta} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \circ \begin{pmatrix} \mathbf{\Sigma}_{11.2} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

The distribution on $\mathbf{\Delta}$ which gives right-invariant measure on $\begin{pmatrix} \mathbf{I} & \mathbf{\Delta} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ is the uniform distribution with element $d\mathbf{\Delta}$. Thus there will be no paradox for $(\mathbf{\Sigma}_{11.2}, \mathbf{\Sigma}_{22})$ for a prior of the form $d\mathbf{\Delta} \,.\, \pi(\mathbf{\Sigma}_{11.2}, \mathbf{\Sigma}_{22})\, d\mathbf{\Sigma}_{11.2}\, d\mathbf{\Sigma}_{22}$. In particular, any prior of the form (A1.1) will, by (A1.2), avoid a paradox, and so yield a marginal posterior for $(\mathbf{\Sigma}_{11.2}, \mathbf{\Sigma}_{22})$ proportional to

$$f(\mathbf{S}_{22}|\mathbf{\Sigma}_{22})\,(d\mathbf{\Sigma}_{22}/|\mathbf{\Sigma}_{22}|^{\frac{1}{2}(v-2p)}) \,.\, f(\mathbf{S}_{11.2}|\mathbf{\Sigma}_{11.2})\,(d\mathbf{\Sigma}_{11.2}/|\mathbf{\Sigma}_{11.2}|^{\frac{1}{2}v}).$$

This product form implies that the marginal element for $\mathbf{\Sigma}_{11.2}$ is proportional to $f(\mathbf{S}_{11.2}|\mathbf{\Sigma}_{11.2})\,(d\mathbf{\Sigma}_{11.2}/|\mathbf{\Sigma}_{11.2}|^{\frac{1}{2}v})$, thus avoiding any paradox for $\mathbf{\Sigma}_{11.2}$.

(iv) *Inference about* $\mathbf{\Phi}_{11.2}$. Since the prior (A1.1) avoids a paradox for $\mathbf{\Sigma}_{11.2}$, we can make inferences about $\mathbf{\Phi}_{11.2}$ from data $\mathbf{S}_{11.2}$ and prior $d\mathbf{\Sigma}_{11.2}/|\mathbf{\Sigma}_{11.2}|^{\frac{1}{2}v}$, where $\mathbf{S}_{11.2}$ has distribution $W(\mathbf{\Sigma}_{11.2}, \nu-q, p)$. Following the analysis in (ii), it can be seen that a paradox for $\mathbf{\Phi}_{11.2}$ will arise unless $v = p+1$. Consequently, no prior of the form (A1.1) can simultaneously avoid a paradox for both $\mathbf{\Phi}_{22}$ and $\mathbf{\Phi}_{11.2}$.

(v) *Multivariate regression.* If we wish to make inferences about $\mathbf{\Delta}$ and $\mathbf{\Sigma}_{11.2}$ from the data, with prior (A1.1), it is clear from (A1.2) and (A1.3) that we can start from the density

$$f(\mathbf{S}_{11.2}|\mathbf{\Sigma}_{11.2})\, f(\mathbf{D}|\mathbf{S}_{22}; \mathbf{\Sigma}_{11.2}, \mathbf{\Delta}), \quad \text{with prior } d\mathbf{\Delta}\, d\mathbf{\Sigma}_{11.2}/|\mathbf{\Sigma}_{11.2}|^{\frac{1}{2}v}.$$

The likelihood is just that which we should obtain from a multivariate regression problem, in which $(\mathbf{x}_{12}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{\nu 2})$ are fixed regressor variables, and $\mathbf{x}_{i1}$ has distribution $N(\mathbf{\Delta}\mathbf{x}_{i2}, \mathbf{\Sigma}_{11.2})$, independently for different i. This model is covered by Zellner (1971, Chapter 8) in some detail, where the choice $v = p+1$ is recommended. Although this is paradox-free for $\mathbf{\Phi}_{11.2}$ as a whole, it is easily seen not to be so for a principal sub-matrix of $\mathbf{\Phi}_{11.2}$.

APPENDIX 2

The Inevitable Paradox of Example 5

Let $\pi(d\zeta, d\xi)$ and $\pi(d\zeta)$ be supposed prior measures for B_1 and B_2 respectively that give the same posterior distributions for ζ. Write

$$p(d\zeta) = \int_\xi \pi(d\zeta, d\xi) \exp\{-\tfrac{1}{2}n(\zeta^2+\xi^2)\} \Big/ \iint \pi(d\zeta, d\xi) \exp\{-\tfrac{1}{2}n(\zeta^2+\xi^2)\}$$

$$p^*(d\zeta) = \pi(d\zeta) \exp(-\tfrac{1}{2}n\zeta^2) \Big/ \int \pi(d\zeta) \exp(-\tfrac{1}{2}n\zeta^2).$$

The equivalence of B_1 and B_2's posterior distributions for ζ for the case $x_1 = x_2 = 0$ implies that $p = p^*$. For the case $x_2 = 0$, the equivalence then implies that, as a function of ζ for each x_1,

$$p(d\zeta) \int_0^\infty t^{2n-1} \exp\{n\zeta x_1 t - \tfrac{1}{2}(nx_1^2+1)\, t^2\}\, dt$$

$$\propto p(d\zeta) \int_0^\infty t^{2n-2} \exp\{n\zeta x_1 t - \tfrac{1}{2}(nx_1^2+1)\, t^2\}\, dt. \qquad \text{(A2.1)}$$

If there were two values of ζ, ζ' and ζ'', say, for which $p(d\zeta)>0$ at ζ' and ζ'', we would then have

$$M_{x_1}(\zeta') = M_{x_1}(\zeta'') \tag{A2.2}$$

where

$$M_{x_1}(\zeta) = \int_0^\infty t^{2n-1}\exp\{n\zeta x_1 t - \tfrac{1}{2}(nx_1^2+1)t^2\}\,dt \Big/ \int_0^\infty t^{2n-2}\exp\{n\zeta x_1 t - \tfrac{1}{2}(nx_1^2+1)t^2\}\,dt$$

But M_{x_1} is monotone increasing in ζ for $x_1>0$, so that, for this case, (A2.2) provides a contradiction. Hence p must be a degenerate probability distribution, in which case the common posterior distribution for ζ would not depend on the data.

APPENDIX 3

Summary of Theory of Structural Inference

A system, operating under stable conditions, yields observed data x in a space X. There is postulated an unobservable error variable e with values in X and known probability distribution P. It is known that x is produced from e by a one–one transformation θ of X onto X. All that is known about θ is that the set of possible values of θ constitute an exact group G say. The structural model is therefore

$$x = \theta \circ e. \tag{A3.1}$$

The structural distribution π_F, say, for θ is what follows from an insistence that the distribution of e must be conditioned by no more than the logical deduction immediately available from the data, namely, that $e \in G\circ x = \{g\circ x \,|\, g\in G\}$. We have

$$\pi_F(\theta\in A\,|\,x) = P(e\in A^{-1}\circ x \,|\, e\in G\circ x).$$

In the special case when $G = HK$, where H and K are sub-groups of G with $H\cap K$ the identity, and we write $\theta = \phi\tau, \phi\in H, \tau\in K$, then ϕ and τ have a joint structural distribution equivalent to that of θ. The *conditional structural distribution* of τ given ϕ is the structural distribution of τ in the modified model

$$y = \tau \circ e \tag{A3.2}$$

where $y = \phi^{-1}\circ x$.

The *marginal distribution* of ϕ has, by definition, a density given by the quotient of the joint density of (ϕ,τ) and the conditional density of τ given ϕ (as would be required for comparibility with the calculus of probabilities).

In the extension in which e has a probability density element $f(e\colon \lambda)\,d\mu_G(e)$, dependent on an additional unknown quantity λ, Fraser proceeds thus: A *marginal likelihood* for λ is defined as follows. Defining $z = G\circ x$, z indexes the orbits of x under G. It may be verified that the distribution of z induced by e is independent of θ. The likelihood corresponding to this probability distribution is the marginal likelihood for λ, which is used to fill the inferential gap.

The strong Bayesian connections of structural theory are as follows (Fraser, 1961; Bondar, 1972): π_F is the posterior distribution for θ corresponding to a prior for θ that is given by a right invariant measure on G. The marginal likelihood for λ is

$$\int f(\theta^{-1}\circ x\colon \lambda)\,d\nu_G(\theta).$$

DISCUSSION ON THE PAPER BY DR DAWID, PROFESSOR STONE AND DR ZIDEK

Professor D. J. BARTHOLOMEW (London School of Economics): The authors are to be congratulated not only on the content of the paper but also on the clarity of their exposition. Instead of burying their main result in mathematical abstraction they have taken the trouble to initiate the reader into the mysteries of the paradox by a succession of examples of real statistical interest. The ordinary reader is thus able to share the deepening perplexity of B_1 and B_2 as consistency, the fundamental tenet of the Bayesian creed, comes under increasing strain.

The danger of translating the higher mysteries of inference into the common tongue is obvious. Every humble practising statistician will want to take part in the debate and we must not be surprised if, when he does, he dismisses the whole exercise as unimportant. It seems clear that B_1 and B_2 are likely to come to much the same conclusions—especially when they have plenty of data. But that is not the issue which is at stake. The real question, it seems to me, is "can Bayesian inference be made objective?" (I will leave others to deal with the implications for structural inference.) This question is important for all those who wish to see statistical inference founded on something more secure than subjective introspection. This paper therefore takes us to the root of the problem of knowledge.

It may help to re-state the case for improper priors as put by Jeffreys. According to him the "ignorance" prior is a purely formal way of getting the inference procedure started. If such a prior distribution were proper it would be possible to make probability statements about the parameters and these are expressions of knowledge, however imperfect, and not ignorance. By choosing an improper prior our ignorance is expressed by the fact that we cannot make any probability statements about the parameters.

In any situation there will be many improper priors which can be used and the problem is to provide criteria to select one from among the many. Invariance arguments have played a central role here. In the one-parameter case the objective Bayesian approach usually produces methods which are scarcely distinguishable from the standard frequentist methods and this has fostered the view that the foundations are more a matter of personal taste. In the multi-parameter case we obviously have to tread more carefully but I am doubtful whether the final outcome is as clear-cut as the authors suggest.

B_1 and B_2 seem to be overmuch concerned with mutual consistency. B_1 never seems to ask whether his posterior distribution is sensible; he apparently does not mind being led into absurdity as long as B_2 is there to keep him company. Let us therefore apply other criteria to the choice of prior and see whether any of the priors seem "right". One possibility is to examine the frequency properties of the inference procedures. I have argued on previous occasions that, for example, a confidence interval constructed from a posterior distribution should be equally likely to include any possible parameter value if the prior is to be a satisfactory representation of ignorance. I hope that someone will be encouraged to look at some of these examples from this point of view but this would be a major project. Instead I shall apply two other criteria, neither of them new, to some of the examples. They are:

(a) to ask whether there are any observations which could occur for which it is "obvious" what the posterior distribution ought to be; if so only a prior which yields this posterior will be satisfactory;
(b) to ask what is the minimum amount of data needed to give non-trivial information about the parameters of interest; the posterior should be proper for such data but improper for any lesser amount.

Example 1 lends itself to investigation by criterion (a). There it seems easy to specify an observational outcome which would leave us no wiser than when we started. Suppose that $n = 3$, $c = 2$ and that we observe x's such that $z_1 = 1$, $z_2 = 2$, $z_3 = 4$. We know that the mean doubles at one of the two possible points but the data seem to provide equal evidence for the change at *both* points. One might therefore require that the posterior in

this case be the same as the prior. The posterior for three priors is given in the table below with elements $d\eta$ (B_1's first choice), $d\eta/\eta$ (his second choice) and $d\eta/\eta^2$. Only $d\eta/\eta$ comes close to our requirement and hence appears the best of the three considered. A similar conclusion is reached using other values of c not too near 0 or ∞ at which points the intuitive case is, perhaps, less clear. (B_1 and B_2 ought to be worried about the correctness of the model in this case but that is, no doubt, outside their narrow Bayesian brief.)

	Prior element	$d\eta$	$d\eta/\eta$	$d\eta/\eta^2$
Posterior $\propto$	$\pi(1)\times$	$\frac{1}{128}$	$\frac{1}{32}$	$\frac{1}{8}$
	$\pi(2)\times$	$\frac{1}{156}$	$\frac{1}{31{\cdot}25}$	$\frac{1}{6{\cdot}25}$

In Example 3 we may well feel that B_1 was too hasty and that he ought to have thought about alternative parameterizations first, but let that pass. Using criterion (b) it is clear that one observation can tell us nothing about the value of ζ since then z must be ± 1. Two observations, however, should enable us to say something about ζ. B_1's posterior distribution is improper with $n = 1$ and proper with $n = 2$ so we cannot fault him on that score. Suppose, returning to (a), that $z = 0$. B_1 and B_2 can now avoid the paradox and so are happy, but ought they to be? If they use the prior $d\zeta/(1-\zeta^2)^{\frac{1}{2}}$ as B_1 proposes, B_1 will have

$$\pi(\zeta \mid z = 0) \propto (1-\zeta^2)^{\frac{1}{2}(n-2)}$$

but B_2 will have

$$\pi(\zeta \mid z = 0) \propto (1-\zeta^2)^{\frac{1}{2}(n-3)}.$$

Given agreement about the prior they do not agree on what can be said about ζ after observing $z = 0$. This is surely just as much an embarrassment as the original paradox. Even if B_1 uses the joint prior which avoids the marginalization paradox then, whatever value of z is observed, B_1 and B_2 will have different posterior distributions. Although B_1 and B_2 use the same data, B_2 appears to get less information from it about ζ. When $z = 0$ he is the poorer by the equivalent of one observation.

Example 4a takes us deeper into the mire since B_1 and B_2 can now only avoid the paradox if they also agree on whether they are interested in just one ζ or both. However, in struggling to avoid the paradox they seem led to priors which cannot bear examination by our criterion (b). It will help if we generalize the problem slightly by supposing we have n_1 observations from $N(\mu_1, \sigma^2)$ and n_2 from $N(\mu_2, \sigma^2)$. The only change to be noted for our purposes is the replacement of $2n$ by n_1+n_2 as the exponent of ω in the integrand of (1.12). The smallest samples capable of providing information about both ζ_1 and ζ_2 are obviously those with $n_1+n_2 = 3$. We thus require p to be such that (1.12) is proper when $n_1+n_2 = 3$ and improper when $n_1+n_2 = 2$. This requires $p = -1$ which would doubtless have been B_1's choice without the unsettling effect of B_2.

Suppose next that B_1 anticipates B_2's interest in ζ_1 alone and so prepares contingency plans. The minimal information for inference about ζ_1 is clearly $n_1 = 2$, $n_2 = 0$. B_1's posterior for ζ_1 under these circumstances would be

$$\pi(\zeta_1 \mid z_1) \propto \int_0^\infty \omega^{n_1-p-3} \exp -\tfrac{1}{2}\{\omega^2 + n_1(z_1\,\omega - \zeta_1)^2\}\, d\omega.$$

With $n_1 = 2$ this again requires $p = -1$ to satisfy criterion (b). B_1 can thus use the "natural" prior, regardless, but trouble with B_2 is inevitable. Once again it appears that B_2 will be worse off than B_1 by, perhaps, the value of up to two observations but the point requires further investigation. B_2's disadvantage will certainly depend on the observed

values of z_1 and z_2. If $z_1 = z_2 = 0$, B_1 and B_2 will avoid the paradox and agree on the inference whatever value B_1 selects for p. With a uniform prior for ζ_1 and ζ_2 they will have a spherically normal posterior distribution for ζ_1 and ζ_2 centred on $(0, 0)$ and it is intuitively clear that knowledge, or lack of it, about σ is irrelevant in this circumstance.

By bringing other priors into consideration and pointing out disagreement even in the absence of the paradox these remarks may seem to make matters worse. In some cases, as in Example 1, our analysis confirms the other approaches so at least there are some areas of multi-parameter inference where objective Bayesian inference is free of serious criticism. In the other two cases our arguments lead to reasonable priors but the marginalization paradox remains. Our attitude must then depend on whether we regard this as a serious obstacle. In practice it probably is not but it remains a theoretical embarrassment nevertheless. However, it is worth remembering that others have their problems too. Frequentist inference in the presence of nuisance parameters is notoriously difficult and I remain to be convinced that the subjective Bayesian can write down a multivariate proper prior with any degree of conviction.

In Stone and Dawid's *Biometrika* paper, B_1 promised never to use improper priors again. That resolution was short-lived and let us hope that these two blinkered Bayesians will find a way out of their present confusion and make another comeback. Both deserve our sympathy but their creators, the authors of the paper, merit a warm vote of thanks for an original and thought-provoking paper.

Mr A. D. McLaren (University of Glasgow): An eminent Oxford statistician with decidedly mathematical inclinations once remarked to me that he was in favour of Bayesian theory because it made statisticians learn about Haar measure. The authors have certainly demonstrated this today, and in a uniquely entertaining way. I would like to thank them very much on your behalf.

Beyond that what have they demonstrated? I think there is some danger of two different targets falling under the same hail of fire. One target is the idea of expressing ignorance by some fully consistent method and the other is the idea that a prior distribution may be improper. These two ideas are related because ignorance priors tend to be improper in typical situations. I am going to present a paradox concerning the first idea, that one can specify an ignorance prior uncontroversially, and it is a paradox which does not contain improper priors at all.

The subject of orientation statistics has recently come into the literature and there is a very interesting paper by Downs (1972) in which he is concerned with data to do with the electrical activity of the heart. Let us think of his data as being a random sample of orientations in three dimensions, or equivalently a random sample of rotations of a rigid body. Let us take it that our model for this distribution of rotations involves a parameter Θ, which itself is an unknown rotation in three dimensions, representing the centre of the distribution. What should the prior distribution for Θ be? The parameter space is the orthogonal group of real orthogonal 3×3 matrices with determinant $+1$.

This situation might arise in geology or soil mechanics, apart from cardiology which Downs was studying. It may help if we consider now the following fanciful example of the same situation, where Θ is a physical rotation. Imagine going off to explore the surface of the moon; there are no windows in our spacecraft and we are going to navigate by dead reckoning. Inside the cabin we have a model of the moon's surface correctly orientated before we start out. In flight and during the exploration our computers in the spacecraft rotate the model in accordance with changes in relative orientation, so that once on the moon the model shows our position and heading. Suppose something goes wrong, and when we get to the moon we find that someone has rotated the sphere inside the spacecraft accidentally. That is Θ, the accidental rotation. Now that we are on the moon we can make some observations; it does not really matter what kind they are, but they will be relevant to Θ. In principle we may have Downs's kind of problem. What

prior distribution is to represent ignorance about Θ? It would seem reasonable that the axis of rotation should be uniformly distributed over the sphere independently of the angle of rotation $\Lambda(\Theta)$ and that Λ should have a uniform distribution on $[0, \pi]$. Would that seem reasonable?

The snag is that if we demand instead the (unique) Haar measure on the rotations in three dimensions, the orthogonal matrices with determinant $+1$, it turns out that the axis of rotation is uniformly distributed over the sphere and is independent of Λ, but unfortunately Λ does *not* have a uniform distribution over $[0, \pi]$. The details are to be found in Miles (1965). I shall at least show you now that Λ cannot have a uniform distribution if Θ has the Haar measure. The real orthogonal matrix Θ has eigenvalues 1, $\exp i\Lambda$, $\exp -i\Lambda$. Thus

$$\text{trace } \Theta = 1 + 2\cos\Lambda.$$

We shall see that, with Haar measure, E trace $\Theta = 0$, implying $E\cos\Lambda = -\frac{1}{2}$. This is in accordance with Miles's density $(2/\pi)\sin^2\frac{1}{2}\lambda$, and certainly shows that the distribution of Λ, is not uniform over $[0, \pi]$.

We find E trace $\Theta = 0$ as follows. trace $\Theta = \Theta_{11} + \Theta_{22} + \Theta_{33}$, the sum of the diagonal elements. By symmetry, each Θ_{ii} has the same distribution; so the trace has expectation three times that of Θ_{11}. But Θ_{11} is the component in a fixed direction of a randomly rotated unit vector, and its distribution is therefore symmetrical about zero, and has expectation zero.

This paradox appears to dispose of the idea that one can specify an ignorance prior uncontroversially. Improper priors are not involved in the paradox because the orthogonal group is compact and therefore has Haar measure (the only relatively invariant measure) that is proper.

Now let us see the relationship of this paradox with the kind of paradox which the authors have been telling us about: it cannot be of the authors' type exactly, because they point out at (1.23) that their kind of paradox cannot happen when the prior distributions are proper. But it does share a common mathematical root, as they put it (Section 5), because the Haar measure for the orthogonal group does not induce the Haar measure for a sub-group comprising the rotations about a given axis.

Passing to improper prior distributions, I think we must accept the authors' paradoxes as demonstrating an important shortcoming of improper priors, where $\pi(\theta)$ does not admit of any marginal distribution for ζ. I do not know how serious this really is in practice, but it is certainly a theoretical limitation. The best way to approach it may be as Professor Stone has shown us in previous papers. If you are worried about an improper prior, then you see if you can approximate it appropriately by a sequence of proper priors.

We have been concentrating today, and I think also over the last few years, on one side of Bayesian ideas—the aim of providing a sound theoretical foundation for inference. We should not forget that there is another side to Bayesian inference: one hopes to get in non-standard situations, particularly, insight from a Bayesian approach that one does not get in any other way.

I might mention as a typical example the calibration problem, where you have bivariate data and observations of X's for fixed Y's, let us say; possibly a linear regression, $X = \alpha + \beta Y + \text{error}$. The only regression you can estimate correctly is X in terms of Y. In future you want to predict Y from the value of X. Those people who are impatient with the complications of Bayesian inference, as indicated by the discussion we have had tonight, should think about this kind of problem. Can one say anything about Y in this situation and if so how can one possibly do it without some prior distribution for Y?

To end on a more positive note, if we are obliged to accept the authors' paradoxes, and if they cannot be explained away by approximation theorems, it is clear we shall have to pay a great deal more attention than in the past to the matter of robustness of Bayesian methods with respect to the choice of a prior distribution, and I would like to mention a

couple of references where such a study seems to be starting. With data $\mathbf{X}$ and parameters $\boldsymbol{\theta}$ suppose (echoing the authors' (1.20))

(a) $E[Y(\mathbf{X}) \mid \boldsymbol{\theta}] = \mu(\boldsymbol{\theta})$,

(b) $E[\mu(\boldsymbol{\theta} \mid \mathbf{X}] = c + dY(\mathbf{X})$.

We have an unbiased estimator Y of a parameter of interest μ and the posterior expectation of μ is a linear function of the unbiased statistic Y. According to Ericson (1969) various things follow, notably $1/v_1 = 1/v_0 + 1/v$. Here v_0 is the prior variance of μ and v is the prior expectation of the sampling variance (given $\boldsymbol{\theta}$) of Y. The posterior variance of μ is a function of $\mathbf{X}$ and v_1 is its prior expectation. This is a very simple example of a result you can get concerning the moments of posterior distributions from light assumptions about the prior. There is also a more interesting theorem by Finucan (1971). Even if assumption (b) above is dropped one can still conclude that $1/v_1 \geqslant 1/v_0 + 1/v$. So useful statements are possible using Bayesian ideas, but not assuming much about the prior distribution, perhaps just the values of the prior mean and variance. Incidentally the Cramér–Rao inequality follows as a simple corollary from Finucan's result.

I should like to second the vote of thanks to the authors for a very stimulating and excellently presented paper.

The vote of thanks was passed by acclamation.

Professor D. V. Lindley (University College London): Many statisticians, myself included, have used "improper" probability distributions for parameters. There were a variety of reasons for this: a feeling that they were convenient approximations to a more complicated specification; their invariance properties; the suggestion that, in some sense, they described a position of ignorance and therefore let the data "speak for themselves"; and, perhaps most importantly, because the procedures that resulted were in good agreement with the "standard" statistical techniques that had stood the test of time. Clearly after tonight's important paper, we should use them no longer. The paradoxes displayed here are too serious to be ignored and impropriety must go. Let me personally retract the ideas contained in my own book. A book that was written as a serious attempt to justify, within the Bayesian framework, much of conventional statistical wisdom. In fact, the Bayesian argument is strong enough to stand on its own feet, and one lesson from this paper is that we should think seriously about our parameter distributions, thinking of θ not as a parameter but as a physically meaningful quantity about which we know something. Let us consider how to express this knowledge, rather than embrace impropriety.

One point that is perhaps not brought out too clearly in the paper concerns the effect of the results on sampling-theory statistics. Quite irrespective of whether or not one accepts the Bayesian viewpoint, it is true surely that the only sensible solutions to a statistical problem are the admissible ones—except that approximations thereto would be acceptable. Now the admissible solutions are essentially the Bayesian solutions. Consequently all standard statistical results must correspond to some Bayesian solution. In practice most of them agree with an improper one, and hence the authors' results must mean that standard statistical ideas must go. Indeed the consequences for the sampling-theory school are more serious than for the Bayesian. For the latter can, and should, take refuge in proper distributions, whereas the former has no refuge from the triumvirate's storm.

Another comment concerns the lazy Bayesian, B_2, who uses only z. In a sense he is not really a Bayesian at all. For suppose the scientist reports data y, z and, with help from B_1, calculates the likelihood function. Then B_1 knows that his inferences depend only on the likelihood function and they would be the same for a second scientist with data that gave the same function. But when B_2 comes along he needs to find the distribution of z,

and he cannot do this from the likelihood function. He has to go back to the scientist and discover the other values of y that might have been observed but were not. He may then integrate over them and find the z-distribution. With two different scientists these might well not agree. Consequently it might appear that B_2 is violating the likelihood principle. When all distributions are proper, the argument that leads to (1.23) shows that, whatever the other y-values were, the result for B_2 would have been the same. Hence B_2 does not really need to enquire of the scientist, he can take any set. Of course, this is not true when improper distributions are used, and the paradoxes remain.

This is an important paper because it clears away so much rubbish from the statistical scene. Let us hope that the cleared highway will encourage more people to think constructively about the parameters and use an honest, proper Bayesian argument. In particular I would like to express the hope that specialists in multivariate analysis will take heed of the results and try to put their own house in order.

Professor BRADLEY EFRON (Stanford University): The question of just what constitutes an uninformative prior in a multiparameter situation becomes ever more vexing, helped along now by the authors' very provocative counterexamples. I only hope that readers will not misread this paper as saying that all is well as long as improper priors are avoided. Suppose we have 100 unknown parameters $\theta_1, \theta_2, \ldots, \theta_{100}$ and data $x_1, x_2, \ldots, x_{100}$, where $x_i \sim N(\theta_i, 1)$, independently given the $\{\theta_i\}$. We may try to represent our lack of prior information on the θ_i by giving them independent $N(0, A)$ priors where A is enormous, say 10^{1000}. This looks uninformative enough, being virtually equivalent to a uniform prior over E^{100} for most purposes, but like the uniform prior it is actually much too informative in some ways.

For example, suppose we wish to estimate $\xi = \sum_1^{100} \theta_i^2$, and observe that the 100 x_i values have sum of squares 200. The *a posteriori* mean of ξ given the data are almost exactly 300 in this case, as opposed to the much more reasonable unbiased estimate $\hat{\xi} = 100$, which has estimated standard deviation 25. Our "uninformative" prior has completely overwhelmed the considerable amount of information in the data! This is because it gives ξ a marginal prior density proportional to ξ^{49} (to a close approximation, for $\xi < 10^{999}$), which is heavily weighted against small values of ξ.

We can correct this by giving A itself a diffuse prior, say with density proportional to $(A+1)^{-2}$, instead of a large fixed value, in which case ξ will have marginal prior density approximately proportional to $(\xi+100)^{-2}$, and the *a posteriori* mean of ξ will always be close to the m.l.e. or to the unbiased estimate. Unfortunately this new uninformative prior is quite informative in its own right. For example, if we wish to estimate $\mu = \max\{\theta_i\}$ and observe $x_1, x_2, \ldots, x_{99}$ to have nearly a $N(0, 1)$ histogram while $x_{100} = 10$, then the *a posteriori* expectation of μ will be close to 5. It is obvious that 10 is a much more sensible estimate in this case.

Why are statisticians interested in uninformative priors? Because they connect Bayesian and frequentist methods, because they offer an "objective" form of Bayesian theory and because they are so convenient for dealing with complicated situations, particularly those involving nuisance parameters. In the 100 parameter problem for instance, a truly uninformative prior, if it existed, would in principle provide a sensible answer to *every* question one could ask about the parameters, both before and after the data were observed. It is worth looking for such a powerful weapon, but sobering to have pointed out that even in much simpler situations the proposed candidates have undesirable properties.

Professor J. DICKEY (State University of New York at Buffalo): Since coherent personal inference implies the use of Bayes's theorem with personal probabilities, and since scientific reporting from statistical data requires objectivity, then a scientific-report-writer should give the posterior probabilities with a variety of prior distributions, typical or bounding of the report-readers' personal uncertainties. Hence, magic unique prior distributions are

without interest, except on occasions when, in practical senses, they approximate real persons' uncertainties.

Constant prior densities and other non-integrable versions of prior "complete ignorance" have long been discredited by, for example, their vulnerability to changes of the parameter-variable, invariance arguments notwithstanding. Such a prior density is prejudiced toward extreme values of the variable, fatally so in high dimensions. Its only legitimate use can be in providing an approximate posterior distribution, when the quality of the approximations is assessed posterior to a realized likelihood function, as by Savage's (1963) system of inequalities named "precise measurement", or "stable estimation". But the constant prior density is in no way unique in the mathematics of "precise measurement", which trivially extends to an arbitrary choice of approximating prior density, even a generalized density.

"Precise measurement" approximates more readily in lower dimensions. To lower the dimensionality (desirable generally), the full-dimensional likelihood function $f(x \mid \theta)$ of $\theta = (\zeta, \xi)$ can be reduced to a personal *weighted* (or *integrated*) *likelihood function* $p(x \mid \zeta)$ of the parameter of interest ζ,

$$p(x \mid \zeta) = \int f(x \mid \theta)\, p(\xi \mid \zeta)\, dQ(\xi),$$

based on $p(\xi \mid \zeta)$ a conditional prior distribution of the nuisance parameter ξ given ζ. In practice, one should use a variety of realistic prior distributions of ξ conditional on ζ. Bayes's theorem applies to the weighted likelihood function,

$$p(\zeta \mid x) \propto p(x \mid \zeta) \cdot p(\zeta).$$

The paper under discussion today warns against lazy or mystical reliance on unrealistic improper prior conditional densities $p(\xi \mid \zeta)$. We have in it yet another nail to seal the fate of *lazy-types* of Bayesianism. How long will it take until the final straw?

The authors treat the factorization (ii), (5.1), for $x = (z, y)$,

$$f(x \mid \theta) = f(z \mid \zeta) \cdot f(y \mid z;\ \zeta, \xi).$$

They repeat the old request for a definition of "no available information concerning ζ in the second factor in the absence of knowledge about ξ". Note that the weighted likelihood function will take the factored form,

$$p(x \mid \zeta) = f(z \mid \zeta) \cdot p(y \mid z, \zeta),$$

where the new personally weighted second factor,

$$p(y \mid z, \zeta) = \int f(y \mid z;\ \zeta, \xi)\, p(\xi \mid \zeta)\, dQ(\xi),$$

in which we have substituted

$$p(\xi \mid z, \zeta) = p(z \mid \xi, \zeta)\, p(\xi \mid \zeta) / p(z \mid \zeta) = f(z \mid \zeta)\, p(\xi \mid \zeta) / f(z \mid \zeta) = p(\xi \mid \zeta).$$

Now, if for realistic specific prior uncertainties $p(\xi \mid \zeta)$, and specific observed y, z, we have the integral $p(y \mid z, \zeta)$ approximately constant in ζ, then there is little available information concerning ζ in the second factor for the given $p(\xi \mid \zeta)$, for example some $p(\xi \mid \zeta) \equiv p(\xi)$.

I would like to draw here the distinction between mathematical invariance and inferential invariance, given a parametrized statistical sampling model. Mathematical invariance refers to a black box labelled "Procedure A" to which a monkey, a man, a robot or other unthinking device inputs observed measurements and then receives statistics computed by the black box. A procedure is mathematically invariant if, without any input information about the scale of measurement, the two output statistics, computed from two input measurements differing only in their scales, will refer with identical implications to the correspondingly scaled unknown parameters.

On the other hand, inferential invariance refers to a statistician's obtaining of coherent personal posterior probabilities for scientists' personal prior uncertainties. The inference

should not vary with known scales with which the measurements are supplied to the statistician. From two scale-differing, but otherwise identical, measurements, the posterior distributions of the two correspondingly scaled parameters will coincide under the change of variable, provided that the two prior distributions similarly coincide. Such prior consistency can be hoped for from real-life assessments which take notice of the scales in which uncertain parameters are discussed, but not from unique magic priors, except within special limited transformation groups *ad hoc* to the statistical sampling models. To quote Lindley (1971), "Why should one's [prior] knowledge, or ignorance, of a quantity depend on the experiment being used to determine it?"

Mr G. N. WILKINSON (Rothamsted Experimental Station): The fundamentals of statistical inference lie beneath a sea of mathematics and scientific opinion that is polluted with red herrings, not all spawned by Bayesians of course. If one's adrenalic propensity for seeing red is stimulated by red herrings, one's view of the fundamentals is likely to become even more opaque. I think an unfortunate aspect of past controversies was not so much that other people's view of inference became fogged but that Fisher's did, on some points in the development of fiducial theory, in particular the uniqueness question on which I have commented before (Wilkinson, 1971) and in which the present paper bears.

The authors are to be congratulated on a very constructive job of reducing the pollution. Having now seen the galley proofs and heard the authors' excellent presentation today, I should like to add the word "brilliant" in describing the paper. It will prove to be a milestone in the development of inference theory, and of fiducial theory in particular, just as an important paper by Lindley (1958) should have been. (At least some of us learnt from that to be much more cautious about applying Bayes's theorem.)

Let me re-iterate first a point that I made at the Cambridge Symposium on "R. A. Fisher's Contributions to Inference in Scientific Reasoning", namely that although Bayesians and Fisherians (fiducialists) can often produce numerically the same inferential probability distributions, the reference sets in which the probabilities are verifiable are fundamentally different. Fiducial limits are confidence limits derived from an appropriately conditioned reference set embodying, in general, one or more fiducial inversions for other parameters, but the assignment of probability at each step is not the same as Bayesian conditioning.

Fiducial theory also gives rise to apparent marginalization paradoxes. The Creasy–Fieller paradox (Creasy, 1954) is an illuminating case. A. T. James, W. Venables and I hope to be publishing shortly, a new fiducial solution which we believe to be the correct one in the relevant context, namely, a fiducial distribution for the angle θ determined by the ratio of means (μ_1, μ_2) of two normal variates (x_1, x_2) with known dispersion matrix, here assumed to be the unit matrix. The solution involves conditioning on the non-centrality statistic $d^2 = x_1^2 + x_2^2$ with a fiducial inversion for the corresponding parameter $\delta^2 = \mu_1^2 + \mu_2^2$ based on the non-central χ_2^2 distribution. It clearly cannot be derived by marginalization from Creasy's bivariate fiducial distribution for (μ_1, μ_2).

A brief comparison of the solution with that of Fieller is set out below for the angle θ measured from the observed angle arctan (x_1/x_2). The solution has a finite condensation

	Upper 95% limits (degrees)	
d^2 (*significance P*)	*New solution*	*Fieller solution*
5·99 (5%)	90·0	53·2
12 (0·025%)	37·3	34·5
24 (0·00006%)	24·2	23·6

of probability associated with the indeterminate angle defined by $(\mu_1, \mu_2) = (0, 0)$. This finite probability depends on the d^2 statistic and is precisely equal to the upper tail

significance probability P for d^2 from the central χ^2_2 distribution. If it is an *a priori* datum of the problem that there is a finite (unknown) probability that $(\mu_1, \mu_2) = (0, 0)$, then this is estimated by the fiducial condensation. If, on the other hand, the *a priori* probability of $(0, 0)$ is known to be zero then the fiducial distribution is interpreted as partly indeterminate with unassignable probability P, and the nominal probability 0·95 say, associated with particular fiducial limits is a lower bound, the upper bound being $(0{\cdot}95 + P)$.

I think the reason why Fieller (and Fisher) overlooked this solution is that they assumed that the existence of the Fieller pivotal statistic precluded the need for any further conditioning, as is the case with the pivotal t-statistic $(\bar{x} - \mu)/s$, where conditioning on s with a fiducial inversion for σ makes no apparent difference to the fiducial solution for μ).

The above solution has thus, from the fiducial point of view, a perfectly acceptable frequency interpretation, but is not, I believe, a conceivable Bayesian *a posteriori* distribution.

The final point I wish to make is that so-called marginalization paradoxes are not paradoxes at all from the fiducial point of view. They correspond to essential inferential distinctions (Wilkinson, 1971) and are paradoxes only within the context of a Bayesian theory of inference. The paradoxes certainly should not and, as the authors have shown, mostly cannot be reconciled by choice of (improper) priors.

[*Added in writing, after the meeting*: Since speaking at the meeting I have found a resolution of some extant logical difficulties with fiducial inference, which I shall describe in more detail elsewhere. The central point is that inference distributions relate fundamentally, not to specific analytical parameters such as $(\sigma_1, \sigma_2, \rho)$, but to the invariants of various transformation groups on the parameter space which carry the fundamental scientific meaning. For instance the one analytical parameter σ_1 may refer to the scale of variation of an x_1-variate alone or, in a parameterization such as $(\sigma_1, \sigma_1/\sigma_2, \rho)$, to the scale or size of the dispersion ellipse of the joint variation of x_1 and x_2, for which a different inference distribution will be appropriate.]

The following contributions were received in writing, after the meeting:

Professor A. P. Dempster (Harvard University): The examples and theory presented by Dawid, Stone and Zidek in their finely crafted paper provide valuable insights into the strange ways of improper prior distributions. While reading, however, I was continually nagged by doubts that the marginalization paradox was real. The key to my trouble may be found in the authors' remark that the position of B_2 is "clear except to the most committed Bayesian". When I am a Bayesian, I am a committed Bayesian, so let me defend the position of B_1 by arguing that B_2 can fairly be said not to exist.

At the close of Section 1, the authors show that the paradox does not arise if the prior element $\pi(d\eta, d\zeta)$ is proper. Actually, their argument requires only that

$$\pi(d\zeta) = \int_{\eta} \pi(d\eta, d\zeta)$$

be finite for almost all ζ. Here lies the nub of the matter, for the paradox rests on the contention that B_2 should be able to duplicate B_1's posterior by combining the likelihood from z with the prior element $\pi(d\zeta)$. If $\pi(d\zeta)$ does not exist, how can there be a paradox? To make the point more concretely, consider the probability element $\pi(d\eta, d\zeta) = d\eta\, d\zeta$ where η and ζ are real-valued parameters. It might appear, as the authors assume, that the associated marginal probability element is $\pi(d\zeta) = d\zeta$. Suppose, however, that the parameters (η, ζ) are replaced by the equivalent pair (ζ, η^*) where $\eta^* = \zeta\eta$. By the usual transformation rules, the probability element $d\eta\, d\zeta$ is expressible in the new co-ordinates as $|\, \zeta \,|\, d\eta^*\, d\zeta$, from which it would appear that the marginal element of ζ is $|\, \zeta \,|\, d\zeta$. The supposed paradox may thus be made to appear and disappear at the whim of an arbitrary choice of co-ordinate system. Another way to phrase the moral of my tale is to remark that the improper Bayesian is in fact proper in his handling of improper integrals, treating

them as limits of any reasonable sequence of proper integrals, whereas B_2 is improper in this sense because he can obtain a wide range of $\pi(d\zeta)$ by carefully selecting different sequences of η regions for different ζ. My point is not to defend improper Bayesians, for I agree that choosing realistic priors is the real game, but improper priors can sometimes be acceptable as limiting approximations to acceptable if not entirely realistic priors.

The parts of the paper dealing with groups and structural inference are deeper and more original. It should be left to Fraser to provide the definitive resolution of the structural paradox, but I would imagine that the committed structuralist would quite happily absorb the new information about the failure of the technician to adjust the knob and proceed accordingly, with a prayer of thanks that the knob fits so nicely into a rotation group.

Dr D. V. HINKLEY (Imperial College): This excellent paper provides an excuse for yet again discussing the various views of probability, in particular as they relate to the question of prior ignorance. Suppose that we try to represent ignorance about a parameter Θ which is somewhere on the real line. One approximation to the appropriate prior distribution is the uniform distribution on $(-A, B)$ where A and B are suitably enormous. This is a *vague* prior, but *informative* because $P(\Theta > 0) = B/(A+B)$. Most proper priors used in practice to approximate prior ignorance implicitly involve an exact relationship between our A and B, so that $\lim_{A,B\to\infty} B/(A+B)$ exists; cf. the abundance of normal priors. But to be *ignorant* about Θ means, in our example, that we know only that $B/(A+B)$ is somewhere in the interval (0, 1). In this context the rotation example of Dr McLaren has nothing to do with prior ignorance. We must decide between ignorance priors and vague priors. My suspicion is that ignorance priors are rarely required because the origin (on whatever scale is appropriate for Θ) usually has physical significance. As Professor Efron has pointed out, these questions become crucial in high dimensions. In practice it would seem necessary to have at least two stages in the prior formulation if anything like ignorance is to be represented. Thus, for example, given a totally new situation in which Θ is two-valued with parameter space $\{\theta_1, \theta_2\}$, we may specify $P(\Theta = \theta_1) = \frac{1}{2}$ if forced to put a number on the probability. We may be led to this by the usual simplified betting arguments, when in fact we really have as prior conviction only that $P(\Theta = \theta_1)$ has some symmetric distribution on the interval (0, 1).

Professor M. R. NOVICK (University of Iowa): The suggestion made by Dawid, Stone and Zidek that statisticians "turn their attentions to the characterization of prior knowledge rather than prior ignorance" is one with which I am in complete agreement. Specifically, in reply to a question following the presentation of my cited paper (Novick, 1969), I definitely rejected the use of improper priors in practice or even proper priors approximating improper ones. My own work on a system for characterizing ignorance had, for me, the sole *practical* outcome of providing a base upon which to construct a proper characterization of prior *knowledge*. This work provided, for me, some justification for the evaluation of specific natural conjugate prior distributions as if all prior information could be equated to a hypothetical prior sample of a certain size. The evaluation of this hypothetical prior sample number can be a useful device in quantifying prior beliefs, as has been demonstrated by Winkler (1967). It is, of course, still true that "improper priors are in widespread current use among Bayesian statisticians", and for that reason alone, the present paper is most welcome.

The final sentence in the Introduction of the present paper is a curious one and requires some comment. The cited paper by Lindley and Smith (1972) deals with problems of simultaneous estimation of many parameters for which a hierarchical model used by Kelley (1927) is adopted, given a firm mathematical and philosophical basis and extended nearly to the point of application to concrete problems. This paper is a monumental contribution to the formalization of a coherent theory of simultaneous estimation,

providing, as it does, a method of incorporating *collateral* information about each of the parameters being estimated, but it does *not* completely solve the problem of specifying *prior* information. Specifically, with this approach, in the final stage of the hierarchy of Bayes's distributions there is still required a subjective specification of a distribution on the parameters of the distribution of the parameters to be estimated. One of the points made in my cited paper (Novick, 1969) is that an ignorance prior *cannot* be used in this model for the between group variance. In a later paper (Novick *et al.*, 1973), I have provided a method for subjectively evaluating a parameter of this type, using the hypothetical prior sample technique to provide a proper prior distribution.

If one is seriously discussing the characterization of prior knowledge one *must* cite the work of Robert Schlaifer (Schlaifer, 1971) and his remarkable MANECON collection of interactive computer programs. I suspect that the further refinement and development of such computer methods will have a far greater effect on the practice of Bayesian statistics than theoretical discussions of ignorance priors or paradoxes arising from their use.

Professor SEYMOUR GEISSER (University of Minnesota): With considerable subtlety and panache, the authors remind us again that improper Bayesian inference does not enjoy the same logical status possessed by proper Bayesian inference. This is not overly surprising as it is well known that quasi-priors could lead to inadmissible estimators and incoherence.

The quasi-Bayesian approach has its source in two primitive data principles. Crudely stated they are: (a) letting the data speak for themselves when little is known or assumed beforehand; (b) the actual units in which you choose to express your observations and appropriate parameter transformations should by and large not effect the inference. The first is often translated into the slack notion that the initial distribution should minimally influence the posterior relative to the likelihood, which in a strict sense already compromises the source. Invariance under suitable transformations is an expression of the second. Simultaneous implementation of both these desiderata often requires dependence on the likelihood and the convenient use of improper priors as a reference. Hence the arch-canon of Subjectivism "A proper prior that in no way depends on the structure of the experiment", is clearly transgressed. Strict interpretation bars proper conjugate priors as well! The fact that the logical implications of Probability could also be violated was already in evidence in the paper by Geisser and Cornfield (1963, p. 373). Consideration of the bivariate normal case and the class of priors $|\Sigma|^{v/2}\, d\Sigma^{-1}$ revealed that only for $v = 3$, would the posterior distribution of the correlation coefficient behave "properly". In early 1964, in an exchange of letters with Professor Stone, I communicated to him a concern relating to partial and simple correlation coefficients even for the "best" candidate $v = p+1$. All this, of course, was barely scratching the surface of the iceberg which he and his colleagues have artfully plumbed.

Does all this mean that a quasi-Bayesian approach is not at all viable? I think not, as long as one is not deluded into believing that it is an exact solution to the inference problem completely consistent with standard probability statements. Within the latter context it will serve as a reasonable approximation for large sample sizes. This the authors themselves aver. For more modest sample sizes, caution was always required, and even more so now until the extent and effect of inconsistency is determined in particular problems. Certainly the same caution is advisable in the strictly subjective framework, lest one be carried away by sheer prejudicial bias, consistency notwithstanding. Even subjectivists find it difficult to completely avoid the use of quasi-priors, though it may be only for the hyperparameters of a proper initial distribution, Lindley and Smith (1972), and thus be a step removed. Whether the quasi-Bayesian approach can meaningfully be embedded in a different "probability" matrix is another matter. The Fiducial and Structural theories seem to have been such attempts, though the founders would probably deny it. In any event, I would hope that the authors not yet take too seriously their own advice and drop everything forthwith for the sake of characterizing prior subjective

knowledge. I would prefer they devote more effort to the development of consistency (less pejorative) indices. For this task, their expertise is abundantly manifest. The exploration of whatever problems ensue for a quasi-Bayesian approach in the prediction of future observations is even more relevant. Too much importance has been vested in parameters and not enough attention paid to the prediction of observables. Here empirical methods can also serve. The investigator has the opportunity to either withhold a fraction of his observations or to secure new ones under the same conditions. Predictions about observables are then subject to public scrutiny in contrast to armchair speculation about hypothetical parameters that may exist only in the models of statisticians.

In conclusion, let me congratulate the authors on a genuinely elegant piece of work.

Professor D. A. S. FRASER (University of Toronto): The opportunity is most welcome to comment on the long, detailed and comprehensive paper by Dr Dawid, Professor Stone and Dr Zidek. Indeed the opportunities do not come easily to comment on the Bayesian viewpoint.

The paper presents two paradoxes, a large collection of examples and certain conclusions concerning statistical inference.

The "paradoxes" or anomalies are not new and have had substantial discussion elsewhere. The examples appear in puzzling profuseness suggesting the precise mechanisms to be illustrated have not been isolated. And the conclusions are unfounded or erroneous.

1. *The paradoxes*

The first "paradox" or anomaly is that a posterior integrated with respect to one parameter need not be a posterior from the component variable on which it is based. Or, in less Bayesian language, it is that a likelihood integrated with respect to one parameter need not be a likelihood or modulated likelihood from the component variable on which it is based (that is, a marginal likelihood or modulated marginal likelihood). From a non-Bayesian viewpoint there does not seem to be anything contradictory or paradoxical in this anomaly. For the common Bayesian practice of representing an unknown by a *measure* on a *space* certainly runs counter to the scientific method of representing an unknown by a *space* (the space of possible values); and of course there is the non-uniqueness of the measure: thus the foundations are such that one can hardly hope for reasonable consequences. What is surprising perhaps is that integration can work at all.

The essentials connected with the first anomaly have been discussed in two papers, Fraser (1971, 1972b).

The first of these papers notes that an integrated likelihood can give anomalous results and that in certain cases no choice of prior can give the appropriate marginal likelihood. Somewhat earlier Professor Lindley (1969) in his *Biometrika* review of *The Structure of Inference* had expressed confidence in this Bayesian integration procedure: "It is integration with respect to this prior that provides the marginal likelihood available for inferences (concerning the second parameter) . . .". The paper was prepared as a rejoinder commenting on this and other inaccuracies in the *Biometrika* review, and was discussed and published as part of the Waterloo Symposium on Statistical Inference. It is of interest that the rejoinder was declined by *Biometrika* with reasons that remain unfathomable; and of parallel interest that the *Bulletin of the American Mathematical Society* solicited a rejoinder to a Bayesian review of the same book and published it with the review.

The second paper (Fraser, 1972b) discusses positive aspects connected with the anomaly —the determination of integration (and other) procedures that do give the appropriate marginal likelihood. It is of interest that this paper too had earlier contact with *Biometrika* and encountered a deeply committed Bayesian referee who reacted vehemently to second round changes made on the basis of his first round recommendations. Indeed it is not easy to comment on things bearing on the Bayesian viewpoint.

The second "paradox" or anomaly is how "two rights can make a wrong"; a better description is "two wrongs do not make a right". In certain contexts the first anomaly can be avoided by using a right invariant prior. The second anomaly is that a right invariant prior on a group may be inconsistent with a right invariant prior on a larger (containing) group, with the consequent recurrence of the first kind of anomaly even with a right invariant prior. As noted above and in Fraser (1972a, 1973a) a right invariant prior may be a "wrong"; and it is not reasonable to expect two wrongs to be consistent or right.

The second anomaly has been discussed in a paper at the Third International Symposium on Multivariate Analysis in 1972 and published as Fraser (1973b).

This paper discusses the factorization of an invariant measure in terms of invariant measures for complementing subgroups, and it discusses the implications of this for Bayesian and structural inference. For the factorization let θ in G be expressable as $\beta\alpha$ where β, α are in complementing subgroups G_2, G_1 respectively; then

$$d\nu(\theta) = d\nu(\beta\alpha) = d\nu_1(\alpha)\cdot\Delta^{-1}(\beta)\,\Delta_2(\beta)\,d\nu_2(\beta)$$

where ν_i and Δ_i are the right invariant measure and modular function for the group G_i. *Note One*: An integration over β (right-coset integration) with respect to ν is not in general an integration with respect to the right invariant measure ν_2. *Note Two*: An integration over α (left-coset integration) with respect to ν is essentially an integration with respect to the right invariant measure ν_2 but it produces a resultant that has in general the additional factor $\Delta^{-1}(\beta)\,\Delta_2(\beta)$ multiplying the right invariant measure ν_2 for the remaining variable α. These two notes describe the simple mechanisms that provide the underlying explanation for the "paradoxes" and the examples.

2. *The examples*

The authors' paper presents a large number of examples in illustration of the two "paradoxes", and the examples are supplemented by a rather detailed and extensive presentation of a group framework for the examples. All of this stands in rather sharp contrast to the short space needed to describe the "paradoxes". A first reaction is that there are far too many examples. A more considered reaction is that the examples are an attempt at illustration and the precise mechanisms to be illustrated have not been isolated.

The first example considers several priors and shows that the right invariant prior avoids the "paradox".

The next two examples are presented in terms of a right invariant for the full parameter and a "paradox" is then obtained for a component parameter; the "paradox" is avoided by suitably modulating the initial right invariant prior. The author's third Example is not new and the anomaly is well known. For they could quote Sprott's result (for example, as recorded in Fraser, 1964) that the posterior for the correlation coefficient from the progression group is Fisher's correlation fiducial; and it has long been noted that that fiducial is not likelihood based.

The next two examples examine modulated right invariants for the full parameter and find that different modulations can be required for different component parameters.

The five examples present a mixed and rather unclear picture—the right invariant is a somewhat natural prior, but to examine a component parameter can require a modulation for the prior, and different modulations can be required for different components. The authors then propose that the confusion can be avoided by a restriction to *proper* priors. This is a strange proposal as a resolution of the difficulties—for it means in the interesting cases that one cannot eliminate a variable, and hence cannot go to the marginal likelihood. The difficulty vanishes because one chooses not to look at it!

There is in fact a simple group pattern underlying all five examples and the pattern is not revealed by the long and detailed group discussion in the authors' Sections 2 and 3.

See Note One above in connection with Fraser (1973b): the right invariant measure factors as

$$dv(\beta\alpha) = dv_1(\alpha)\cdot\Delta^{-1}(\beta)\,\Delta_2(\beta)\,dv_2(\beta),$$

and a right-coset integration with respect to β is not in general an integration with respect to the right invariant v_2 for the parameter being eliminated—because of the factor $\Delta^{-1}(\beta)\,\Delta_2(\beta)$. All four component-parameter examples involve right-coset integration with respect to a natural group, and the introduction of a modulating factor is seen to be merely a device that eliminates the "extraneous" factor $\Delta^{-1}(\beta)\,\Delta_2(\beta)$. Thus the substance underlying the first "paradox" and the five examples can be found in the right-coset factorization (Note One) of a right invariant measure.

In the framework of a structural model a right-coset integration on the parameter space corresponds to a left-coset integration on the error space; and it is known that such an integration violates the probability principles needed for the conditional distributions. Cautions in this regard were expressed in *The Structure of Inference*, and Dr Stone may recall a slighting remark on one of these cautions in his review of *The Structure of Inference*. The examples show a Bayesian need for an analogue of such cautions.

The next two examples do not have the simple group structure found with the first five examples. Accordingly there is no right invariant measure and it is not surprising that integrated likelihood cannot give the marginal likelihood for the component parameter. Each example can, however, be obtained from a larger model that does have group structure, and indeed obtained by a left-coset integration on the error space. The authors examine some aspects of this larger group model. The underlying mechanisms, however, are not revealed but again they can be found by the use of Note One. For this it is easier to translate from right cosets on the parameter space to the corresponding left cosets on the error space. Two levels of left-coset integration are involved: to obtain the distribution of the given variables; and to obtain the distribution for the marginal variable. No group step exists between the levels and accordingly there is no right invariant measure. Thus as would be expected the "paradox" occurs.

3. *The structural example*

The authors consider the bivariate normal as generated by a rotation and a positive lower triangular transformation from standard normal variables. And they conclude "that the theory of structural inference is powerful enough to develop its own paradoxes without the assistance of improper Bayesians".

The multivariate form of this bivariate model has been examined in detail as the central example in the paper "Inference and redundant parameters" presented at the Third International Symposium on Multivariate Analysis in Dayton, June 1972 and published as Fraser (1973b). Whatever power the "theory" may have—some comment on this later—and contrary to the authors' assertion, no paradoxes or contradictions are involved in the analysis of this bivariate or multivariate structural model. For some comments on difficulties with a weaker model, see Fraser (1973d).

4. *The conclusions*

The authors' "paradoxes" and examples have drawn attention to the difficulties and inconsistencies connected with the use of left invariant or *flat* priors, difficulties that derive as we have seen from factorization properties of the invariant measures. The difficulties provide a moderate range of indeterminacy in the possible posterior distributions. The authors conclude "that more statisticians will be guided by the philosophy of Lindley and Smith (1972) and turn their attentions to the characterization of prior knowledge, rather than prior ignorance".

The question whether subjective priors, personal biases of investigators, should be used in the statistical analysis of data and a statistical model will not be raised here; comments may be found in Fraser (1972a). Rather we consider the range in which an

investigator could reasonably present a subjective prior. All indications are that this range would be far wider than that connected with the indeterminacy of the posterior in the flat Bayesian analysis.

It seems then that the authors, on facing the indeterminacy with flat priors, are in fact making a virtue of indeterminacy by their suggested use of subjective proper priors. Certainly Professor Lindley's use of normal priors leads to nice multivariate calculations and *we* are entitled to use our own priors—good luck with the integrations—but going to greater indeterminacy can hardly be taken seriously as a *conclusion* from the indeterminacies discussed in the authors' paper.

Concerning the structural approach the authors suggest "that the adjacent layers of the theory are in basic conflict". As has been emphasized in Fraser (1972a) the structural approach is primarily concerned with finding those portions of statistical analysis that follow *necessarily* from the data and the statistical model alone. With data and the classical model one obtains the likelihood function and its model. With data and the structural model one obtains a conditional distribution that separates into components corresponding to parameters of interest. Classical *theory* can then provide the small and usually immediate step to tests of significance, point estimates and confidence intervals. The data and the structural model also label possible parameter values onto the space of the conditional distribution just described; perhaps the Bayesians' feeling of exclusive rights to posterior distributions draws their attention to this aspect of structural analysis—after all, all the Bayesian has is posteriors. Up to this point there is no "theory", only the determination of necessary results or logical consequences; the term "structural analysis" seems appropriate to refer to this. The authors almost touch on this aspect in their remark ". . . must be conditioned by no more than the logical deduction immediately available from the data . . .". The problem of determining things that are necessary within statistical analysis is an important one that concerns us all. This is particularly so if we consider the large number of principles and criteria and reduction methods that have been introduced to statistical analysis in the last 30–40 years in order to get results. A reassessment to determine necessary results seems important to me.

The use of the structural model is related to the objective identification of sources of variation or error; see Fraser (1972a). For example, is there an identifiable source for the error affecting the first variable x_1 in the authors' example? It seems easy to treat all events with their corresponding probabilities on an equal basis, but the results of Basu concerning ancillary statistics show that this is premature. Some recent comments on the consequent difficulties for the standard ancillary concept have been discussed in Fraser (1973c).

Concerning the authors' suggestion then, we are not involved with "adjacent layers" of a "theory" but rather with two different statistical models and whether one or the other is the appropriate model in some application.

5. *Some details*

The authors comment on "the expression of ignorance by means of invariance . . . while Fraser's . . . theory of structural inference makes it almost axiomatic". As remarked above most of structural inference is concerned with the *necessary* analysis of the structural and classical models, not with theory. Ignorance is *not* expressed there by means of invariance as the authors suggest; it is expressed by the set, the *set* of possible values for the unknown.

". . . while Fraser's theory is somewhat controversial at the initial axiom level". Structural analysis is involved with a model and the necessary analysis of it. No axioms are involved beyond that needed for the mathematics used to describe the model; for example, to describe a probability space.

It seems unusual to include the proof without references for the standard invariance result: the distribution of the maximal invariant variable depends only on the maximal invariant parameter.

Professor ARNOLD ZELLNER (University of Chicago): As the authors of this interesting paper state, improper (and/or locally uniform†) "ignorance" prior distributions are in widespread use among Bayesian statisticians, including H. Jeffreys, D. V. Lindley, G. E. P. Box, G. C. Tiao, S. Geisser and others. Such priors serve the purpose of allowing mainly the information in the data to be reflected in posterior distributions, a very useful procedure when an investigator is indeed ignorant about parameters' values or wishes to proceed as if he were ignorant. This in no way implies that such priors are always to be used. Obviously use of informative prior distributions can serve extremely useful purposes and some work has been done to formulate and use such distributions (see e.g. Raiffa and Schlaifer, 1961; Zellner and Geisel, 1970; Zellner, 1971, 1972; Zellner and Vandaele, 1973; Zellner and Williams, 1973).

That "no fully acceptable theory" of prior distributions to represent "knowing little" or "ignorance" is available is indeed lamentable and recognized in the work of Jeffreys (1967) and others. However, in many fields of science, it is often the case that practice leads theory and theory serves a useful function in analysing, rationalizing and extending successful practice. For example, in the present paper the authors draw attention to a "marginalization paradox" with which users of improper, ignorance priors should be aware. However, the authors are woefully silent on the *practical* implications of the paradox and on useful methods for circumventing it. On the former point, they do state briefly in their concluding section that, "It is obvious that, in cases involving a large number n of replicate observations, the inconsistency will be relatively unimportant . . . ." No indication is given of how large n must be for this conclusion to hold. Further, as pointed out below, some of the authors' suggested improper priors, that are not subject to the paradox, imply other results that may not be considered entirely satisfactory.

As regards some of the practical implications of the paradox, in the authors' Example 1, they state that when the improper prior for the scale parameter is "in agreement with the usual prescription for a scale parameter", there is no paradox. In Example 2, use of the prior $d\mu_1\, d\mu_2\, d\sigma/\sigma^2$, that avoids the paradox, rather than the "usual" prior, $d\mu_1\, d\mu_2\, d\sigma/\sigma$, results in just a change of one degree of freedom in the marginal posterior for μ_1 and μ_2, hardly a significant change when the sample is even moderately large. Similarly in Example 3, use of $d\gamma\, d\sigma_1\, d\sigma_2/\sigma_2^2$, that avoids the paradox, rather than $d\gamma\, d\sigma_1\, d\sigma_2/\sigma_1\, \sigma_2$, results in only a change of one degree of freedom in the Student-t posterior for γ and minor changes in the posteriors for σ_1 and σ_2. In these examples, as well as others, it would be extremely useful to have these points relating to sensitivity of posteriors more fully investigated, either analytically or by computer calculations using data generated from known models.

It should also be noted that when the improper priors that avoid the paradox are employed in Examples 2 and 3, the Student-t posterior distributions for μ_1 and μ_2 and for γ will not be the same as the related sampling distributions. Thus an "inconsistency" similar to the one mentioned in the second paragraph of the paper is present when the authors' suggested priors that avoid the paradox are employed in practice. It was similar "inconsistencies" that led Jeffreys (1967) to modify his principle of generating non-informative priors according to $|\text{Inf}|^{\frac{1}{2}}$ in the case of several normal means, a case similar to Example 2.

While it is true that use of proper, informative priors avoids the paradox that the authors present, it is obvious that when little or no prior information is available or when an investigator wishes to proceed as if ignorant, the suggestion that informative, proper priors be used cannot be implemented. Further, in an ignorance situation, there is a great danger that use of informative priors may introduce "mis-information" in an analysis.

Finally, it appears that there are usually at least two objectives involved in generating non-informative priors, namely, (1) expressing ignorance and (2) obtaining priors with various invariance properties. Sometimes, as Jeffreys recognizes at many points in his

† While "locally uniform" priors are usually defined to be proper, they are generally used in practice as if they were improper.

work, the second objective may be in conflict with the first. That is, by insisting on broad invariance properties, it may be difficult to achieve the first objective of expressing ignorance adequately. Of course, Jeffreys is not dogmatic in applying his invariant rule, $|\text{Inf}|^{\frac{1}{2}}$, in practice. A possible, tentative approach to this problem involves formulating a principle that achieves the first objective of adequately expressing ignorance in the choice of a prior, a prior that is tailored to the particular parameterization and model for the observations that are employed. For example, let $p(y \mid \theta)$ be the p.d.f. for y and

$$I(\theta) = \int p(y \mid \theta) \log p(y \mid \theta)\, y,$$

the information in the data distribution. Then consider

$$G = \int I(\theta)\, p(\theta)\, d\theta - \int p(\theta) \log p(\theta)\, \mathrm{d}\theta,$$

the average information in the data distribution minus the information in the prior, $p(\theta)$. If G is maximized by varying $p(\theta)$ subject to $\int p(\theta)\, d\theta = 1$, the result is $p(\theta) = k \exp\{I(\theta)\}$, where k is a normalizing constant.† This procedure, described in Zellner (1971, pp. 50–52), has been applied to several problems with encouraging results. For example, if $y = \mu + \epsilon$ with $\epsilon \sim N(0, 1)$, the solution is $p(\mu) \propto \text{const.}$, $-M < \mu < M$. If $y = \mu + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$, $p(\mu, \sigma) \propto 1/\sigma$, $-M < \mu < M$ and $0 < a_1 < \sigma < a_2$. Further, if $\phi_x = \sigma^x$, $x = 2, 3, \ldots$, application of the principle leads to $p(\mu, \phi_x) \propto 1/\phi_x^{1/x}$ which is more spread out than $p(\mu, \phi_x) \propto 1/\phi_x$, the result that Jeffreys gets by insisting on invariance with respect to powers of σ and imposing the condition that μ and ϕ_x be *a priori* independent. In other normal problems and the rectangular, exponential and binomial distributions, similarly encouraging results have been obtained. [For the rectangular p.d.f., $p(y \mid \theta) = 1/\theta$, $0 \leqslant y \leqslant \theta$, $p(\theta) \propto 1/\theta$; for the exponential $p(y \mid \theta) = \theta^{-1} \exp\{-y/\theta\}$, $0 < y < \infty$, $p(\theta) \propto 1/\theta$; and for the binomial, $p(y \mid \theta) = \theta^y (1-\theta)^{1-y}$, $y = 0, 1$, $0 \leqslant \theta \leqslant 1$, $p(\theta) \propto \theta^\theta (1-\theta)^{1-\theta}$. Note that this last proper prior lies between the priors $p(\theta) \propto \text{const.}$ and the priors $p(\theta) \propto \theta^{-a}(1-\theta)^{-a}$, $a = 1$ or $a = \frac{1}{2}$, priors that were cogently discussed by Jeffreys (1967, pp. 123–125). In addition, G can be maximized subject not only to the condition that $p(\theta)$ is proper but also subject to conditions on the first and/or second moments of θ, i.e. $\int \theta p(\theta)\, d\theta = \mu_1$ and $\int \theta^2 p(\theta)\, d\theta = \mu_2$, with μ_1 and μ_2 given constants.]

The authors replied in writing as follows:

Professor Bartholomew goes beyond the ambit of our paper in several interesting ways. With a Jeffreys–Lindley–Novick poker, he stirs the ashes to see if he can raise an objective Phoenix; we wish him all success. However, on matters that bear directly on the content of our paper we have one caveat, one comment and one correction.

With reducibility there are two distinguishable aspects of B_1's posterior distribution: (1) its Bayesianity/unBayesianity when agreement with B_2 is possible/impossible and (ii) its degree of "absurdity". In ignoring "mutual consistency" in his analysis of Example 1, Professor Bartholomew shows that he is prepared to countenance B_1's unBayesianity.

Confidence regions for ζ based on z will not, in general, be invariant with respect to T (see Section 2.4) so that the confidence region property associated with right-invariant prior on T will not apply to ζ. For example, the Creasy interval for ζ_1/ζ_2 is known not to have the property.

Professor Bartholomew's discussion of Example 3 is erroneous; if B_1 uses the prior that avoids the paradox then B_1 and B_2 *will* agree if B_2 uses $d\zeta/(1-\zeta^2)^{\frac{3}{2}}$.

We were fascinated by Mr McLaren's "lunatic" example, which shows dramatically that invariant measures may be unnatural. Incidentally, Jaynes's analysis (1971) of the Bertrand paradox makes the contrary view seem equally attractive. Clearly there is a dependence on context.

† That the solution, called a maximal data information prior, is a proper prior may be satisfying to some. To have $p(\theta)$ proper, it is often necessary to give θ a finite range as in the case of locally uniform priors.

Mr Wilkinson leaves the Fishermen to catch red herrings and, swimming against the current, is soon out of our depth. We wish him luck in his adventures, but prefer to remain behind on the margins of the dry land.

Dr Hinkley and Professor Efron remind us that innocuous-looking prior distributions can have extreme and perhaps surprising implications. Efron's investigation of the statistical consequences of certain priors is very welcome, since it is only by comparing such consequences with intuition that the reasonableness of a prior may be assessed. However, if there is a clash it *could* be that the intuition needs sharpening. Hinkley's attempt to express ignorance at a deeper level than usual is reminiscent of the multi-stage models of Lindley and Smith (1972), but in his example the two approaches would lead to identical analysis.

Professor Novick has faith in the concept of an ignorance prior as a zero level from which knowledge may be measured. This appears to avoid mathematical inconsistency and to be of some help in practical assessment, but there is the danger that the zero level may not be the "fixed point" one would like it to be.

Professor Lindley wonders whether B_2 is a *real* Bayesian. Since a real Bayesian should make inferences conditional on all the data available, he is not; but at least he uses his reduced data in truly Bayesian fashion. The question is somewhat unimportant, since B_2's dramatic function is merely to highlight B_1's inconsistencies. We make no claim that his desire to restrict attention to z alone is wise, only that he *appears* to be following B_1 in this.

Our purpose in Section 1 was to show that B_1's improprietous progress from likelihoods for $\theta, \{L_t(\theta)\}$, to marginal posteriors for ζ could have rakish consequences. Professor Lindley may be quite right to demand that the *force* of this demonstration should not depend on s in any specification of the sampling distribution $p(x \mid \theta) = s(x) L_{t(x)}(\theta)$. We could respond to this demand by calling B_1 "unBayesian" if there is *any* s that leads to the marginalization paradox. In fact, this slight reformulation of our case against B_1 has the consequence of enlarging the area of applicability of our investigation of impropriety.

The introduction of sampling distributions in an analysis of Bayesianity should surprise only those Bayesians who have kicked away the ladder by which they climbed onto the likelihood-cum-prior platform. The axioms of coherent behaviour and the likelihood principle relate to one another the actions or inferences that are relevant in *different* situations; one can only have coherence when there is more than one thing to cohere. Thus it is the essence of the marginalization paradox that all possible data points x should be considered together with their sampling distributions. This is the setting in which one introduces the axioms of coherent behaviour and it is the appropriate setting in which to challenge them.

Grateful as we are for Professor Lindley's generous comments on our "rubbish clearance" operation, we confess to having no jointly coherent view about the slogan "standard statistical ideas must go" and to being far from optimistic about where the "cleared highway", no doubt paved with good intentions, leads.

Professor Dempster has used the propriety analysis at the end of Section 1 as a platform from which to emphasize the near vacuity of the concept of marginalization of an improper prior. On this matter we are in full agreement with him and are puzzled as to why he believes that our paradox has any connection with it. Perhaps it is necessary to reformulate our message a little: *A particular application A of a statistical method M applied to data x leads to inferences about ζ based only on z; the same method M applied to data z will always lead to inferences of a different overall character.* In the examples, M was the use of Bayes's theorem with possibly improper priors while A was determined by B_1's choice of (improper) prior; B_2's choice of prior for use with z was *quite unrestricted.* However, when A corresponds to $\xi \equiv 1$ in (2.10) then, as (2.8) and (2.11) show, B_2 will agree with B_1 if he uses the $d\zeta$ of (2.10). In this connection, reparametrization is quite irrelevant.

The same point arises in connection with Professor Geisser's relevant reference to his own work with Cornfield on the bivariate normal. Their definition of "proper" behaviour of the posterior distribution of the correlation coefficient ζ confounds two issues (i) the

possibility of agreement between B_1 and B_2 when B_2's choice of prior for ζ is unrestricted, (ii) its possibility when B_2's choice is restricted to the "natural" marginal prior for ζ. It is a mathematical accident that these two requirements lead to the same choice.

As Beran (1972) makes clear, Professor Dempster's own "upper and lower probability" theory is a generalization of Fraser's structural theory. Why, then, the appeal to a lower court for the "definitive" judgment on the structural paradoxes to which we have drawn attention?

Professors Fraser and Zellner attribute to us some concrete proposals which we did not in fact intend to make. Thus Fraser says that we "propose that the confusion can be avoided by a restriction to *proper* priors", while Zellner seems to believe that we positively suggest certain improper priors as reasonable. Any constructive comment which might be found in the paper is accidental, since it was our deliberate intention to leave open the question "What *should* one do?". The discussants themselves have provided some interesting answers to this question. We ourselves are not in full agreement on the implications of the paradoxes, but if some sort of merged opinion were attainable and were of interest it might be that the most promising avenue of escape for a Bayesian is that signposted by Professor Dickey in his contribution here, and investigated by him elsewhere. The Bayesian should attempt an assessment of his true prior beliefs for his own satisfaction. For scientific reporting, we may distinguish two cases. If a large number of data are available then some "precise measurement" results might apply, and it may be adequate to pretend that the prior was conveniently improper—even paradoxical. However, the adequacy will vary with the observed data and needs careful investigation. With small numbers of data it appears that attempts to describe ignorance are stillborn, and the most informative summary of the data lies in the presentation of the posteriors derived from a wide range of reasonable—or even extreme—prior distributions.

Presumably a large number of non-Bayesians are happy to base their inferences for ζ on the distribution of z, just like B_2, and many practical Bayesians may be tempted to do the same. Rather than concentrate on finding an "index of consistency" for B_1 it may be more fruitful to investigate the extent to which such an approach might be justified as an approximation.

In *The Structure of Inference*, Professor Fraser carefully avoided any involvement with the Bayesian method. In our analysis of the very simple example of Section 4, we have done the same; moreover we have not departed from the precepts of the book. It is therefore disappointing that Fraser has declined the opportunity to discuss this example in a way that would clarify the grounds for his refutation of our charge of inconsistency. Nine-tenths of his comments clearly refer to the Bayesian part of our work and the structural example is dismissed with a reference to some of his unpublished papers. Where can we find an understanding of the "appropriateness" of structural models that will save us from a fate such as that of *F*, traumatized by the trivial turns of the technicians' tale? It is easy to see why Professor Fraser has to invoke an excathedral term such as "appropriateness". He clearly believes that structural probability is not a bit of speculative methodology but a necessary and unavoidable consequence of data and structural model. In a given problem he therefore cannot admit more than one such model; whereas, in our example, we believe that the different structural modellings of the same error situation should not have different statistical consequences. As to the Bayesian (or "integrated likelihood") nine-tenths, who can quarrel with a Deuteronomy—plus or minus a few minutiae?

We agree, however, that a reference to Sprott would have simplified Example 3. Incidentally, the "maximal invariant" reference is p. 220 of Lehmann's *Testing Statistical Hypotheses* (1959).

References in the Discussion

Beran, R. J. (1972). Upper and lower risks and minimax procedure. *Proc. 6th Berkeley Symp. Math. Statist. Prob.*, **1**, 1–16.

DOWNS, T. D. (1972). Orientation statistics. *Biometrika*, **59**, 665–676.
ERICSON, W. A. (1969). A note on the posterior mean of a population mean. *J. R. Statist. Soc.* B, **31**, 332–334.
FIELLER, E. C. (1956). Some problems in interval estimated. *J. R. Statist. Soc.* B, **18**, 175–185.
FINUCAN, H. M. (1971). Posterior precision for non-normal distributions. *J. R. Statist. Soc.* B, **33**, 95–97.
FRASER, D. A. S. (1964). On the definition of fiducial probability. *Bull. Int. Statist. Inst.*, **40**, 842–856.
—— (1971). Events, information processing and the structured model: Addendum. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds). Toronto: Holt, Rinehart and Winston.
—— (1972a). Bayes, likelihood or structural. *Ann. Math. Statist.*, **43**, 777–790.
—— (1972b). The determination of likelihood and the transformed regression model. *Ann. Math. Statist.*, **43**, 898–916.
—— (1973a). Comparison of inference philosophies. *Theory and Decision*, **4** (in press).
—— (1973b). Inference and redundant parameters. In *Multivariate Analysis, III.* New York: Academic Press.
—— (1973c). The elusive ancillary. In *Proceedings of the Halifax Seminar on Multivariate Statistical Inference.* Amsterdam: North-Holland.
—— (1973d). Comments on the McGilchrist paper. *J. Amer. Statist. Ass.*, **68** (in press).
JAYNES, E. T. (1971). The well-posed problem. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds). Toronto: Holt, Rinehart and Winston.
JEFFREYS, H. (1967). *Theory of Probability*, 3rd revised ed. Oxford: Clarendon.
KELLEY, T. L. (1927). *Interpretation of Educational Measurements.* Yonkers on Hudson, N.Y.: World Books.
LINDLEY, D. V. (1958). Fiducial distributions and Bayes's theorem. *J. R. Statist. Soc.* B, **20**, 102–107.
—— (1969). Review. *Biometrika*, **56**, 453–456.
—— (1971). *Bayesian Statistics: A Review.* Philadelphia: Society of Industrial and Applied Mathematics.
MILES, R. E. (1965). On random rotations in R^3. *Biometrika*, **52**, 636–639.
NOVICK, M. R., LEWIS, C. and JACKSON, P. H. (1973). The estimation of proportions in m-groups. *Psychometrika* (in press).
RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory.* Boston: Graduate School of Business Administration.
SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193–242.
SCHLAIFER, R. (1971). *Computer Programs for Elementary Decision Analysis.* Boston: Graduate School of Business Administration.
WINKLER, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *J. Amer. Statist. Ass.*, **62**, 776–800.
ZELLNER, A. (1972). On assessing informative prior distributions for regression coefficients. Unpublished paper, University of Chicago.
ZELLNER, A. and GEISEL, M. S. (1970). Analysis of distributed lag models with applications to consumption function estimation. *Econometrica*, **38**, 865–888.
ZELLNER, A. and VANDAELE, W. (1973). Bayes–Stein estimators for k-means, regression and simultaneous equation models. In *Studies in Bayesian Econometrics and Statistics: Essays in Honor of Leonard J. Savage* (S. Fienberg and A. Zellner, eds). Amsterdam: North-Holland (in press).
ZELLNER, A. and WILLIAMS, A. D. (1973). Bayesian analysis of the Federal Reserve-MIT-PENN model's Almon lag consumption function. *J. Econometrics* (in press).