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 $\{P_\theta | \theta \in \Theta\}$ 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 $\bar{G} = \{\bar{g} | g \in G\}$ 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

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

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

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

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 $\{z(x) | x \in X\} = Z$, 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 $\mathscr{y}$ 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

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

Proof.

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

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) $f(y,z|\eta,\zeta) = f(y|z;\eta) f(z|\zeta)$ 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 $\boldsymbol{\Sigma}$ of a Wishart variable $\mathbf{S} \sim W(\boldsymbol{\Sigma}, \nu, p)$. 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

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

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

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

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

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

Hence

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

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

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

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

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

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

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

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

Thus (3.10) is proportional to

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

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

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

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

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 $t \in T$,

$$\int_{\bar{T}} f_0(e \mid z;\ \eta, \zeta)\, \delta(\bar{t}\eta, \zeta)\, \mu_{\bar{T}}(d\eta)$$

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

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

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

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

(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).

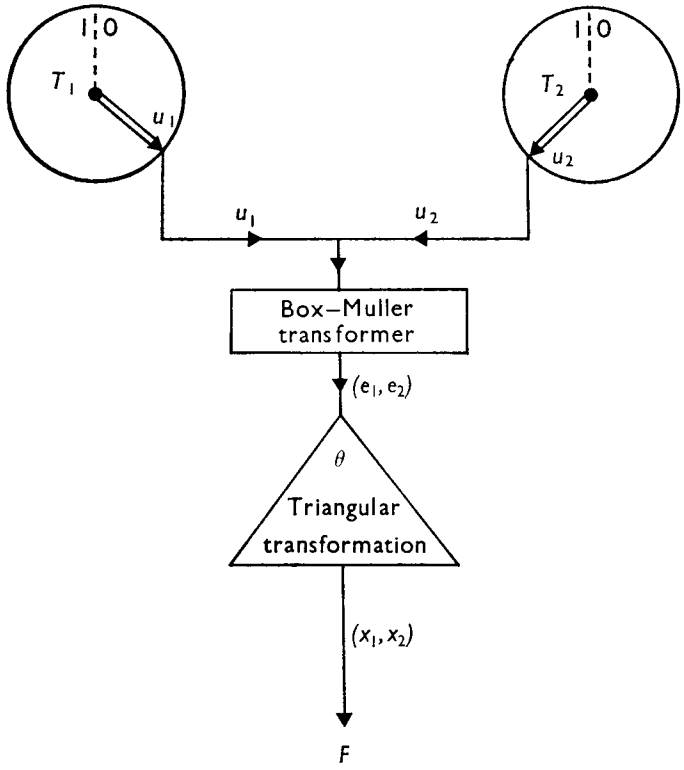

FIG. 1. The bivariate normal generator.

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

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

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

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

(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 $W(\boldsymbol{\Sigma}_{22}, v, q)$, 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 $W(\mathbf{\Sigma}_{11.2}, v-q, p)$. 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 $(\mathbf{x}_{12}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{v2})$ 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

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