INVARIANT PRIOR DISTRIBUTIONS[1]

By J. Hartigan[2]

Princeton University

0. Summary. The paper is mainly concerned with determining prior distributions on ignorance over parameter spaces, using invariance techniques similar to those of decision theory. Prior distributions are rarely determined exactly by such techniques and a number of less compelling methods for exact determination are given.

1. Introduction. Suppose we are observing a random variable X, with probability density $f(x \mid \theta)$, $\theta \varepsilon \Omega$, and we wish to make inferences about θ on the basis of the observed value x of X.

The Bayesian method is to assume some density $h(\theta)$ over Ω to represent our prior knowledge of θ. The posterior distribution of θ given x, obtained by applying Bayes' product formula, has density $g(\theta \mid x)$, where

$$g(\theta \mid x) \propto f(x \mid \theta)h(\theta).$$

A traditional difficulty is deciding on a density $h(\theta)$ to represent our prior knowledge. If we could decide on a prior density to represent ignorance of θ, we could suppose our prior knowledge to be equivalent to some specific set of observations of X, and obtain a density representing the prior knowledge using Bayes' formula.

Our object here is to obtain prior densities on ignorance using invariance techniques.

Jeffreys [1], 1946, appears to have been the first to base selection of a prior density on ignorance on properties of the family $f(x \mid \theta)$. We will follow Jeffreys' practices of allowing prior densities to have $\int_\Omega h(\theta)\, d\theta = \infty$, of leaving all prior densities unnormed, and of identifying prior densities which differ by a constant multiplier. [We will also allow posterior densities the same freedom.]

2. Inversions. Let X be a random variable with density $f(x \mid \theta)$, $x \varepsilon S$, an open subset in R^n, and $\theta \varepsilon \Omega$, an open subset in R^k. Let $f(x \mid \theta)$ have a continuous θ-derivative for all $x \varepsilon S$, $\theta \varepsilon \Omega$. Let F denote the family of densities $f(x \mid \theta)$, $\theta \varepsilon \Omega$.

The posterior density $g(\theta \mid x)$ is fundamental; our interest in prior densities is so that we can determine the posterior densities by the equation $g(\theta \mid x) \propto f(x \mid \theta)h(\theta)$.

We wish to find reasonable posterior densities $g_F(\theta \mid x)$ for each family F. An *inversion* g is any function which assigns a posterior density $g_F(\theta \mid x)$ to

Received 29 January 1963; revised 25 August 1963.

[1] Work supported partly by the Office of Naval Research.

[2] Now at the Statistical Laboratory, Cambridge University.

each F. An *inversion kernel* h is the associated function which assigns a prior density $h_F(\theta)$ to each F and which determines an inversion g through the relation $g_F(\theta \mid x) = f(x \mid \theta)h_F(\theta)$.

Essentially an inversion kernel is a method of assigning prior distributions. It is necessary to consider a method of assigning prior distributions to a class of F's rather than to a single F, because our basic technique is to relate prior distributions for different F's.

3. Desirable properties for inversions.

I. *S-labelling invariance.* Suppose that F and F^* are two families of densities such that there exists a differentiable 1-1 transformation z of S onto S^*, $x \rightarrow zx$, such that

$$f^*(zx \mid \theta)(dzx/dx) = f(x \mid \theta) \qquad \text{for all} \quad x \varepsilon S, \theta \varepsilon \Omega.$$

We say that g is *S-labelling invariant* if for each such transformation z, $g_{F^*}(\theta \mid zx) \propto g_F(\theta \mid x)$.

If we let $X^* = TX$ denote the transformed random variable, Property I ensures that the observation $X = x$ has the same relevance to θ as the observation $TX = Tx$.

II. *Ω-labelling invariance.* Suppose F and F^* are two families of densities such that there exists a differentiable 1-1 transformation T from Ω onto Ω^* such that

$$f^*(x \mid T\theta) = f(x \mid \theta), \qquad \text{all} \quad x \varepsilon S, \theta \varepsilon \Omega.$$

We say that g is *Ω-labelling invariant* if $g_{F^*}(T\theta \mid x)(dT\theta/d\theta) \propto g_F(\theta \mid x)$.

Essentially F and F^* are the same family of distributions, identified by the transformation $\theta \rightarrow T\theta$. Property II ensures that the posterior distributions for F and F^* are also identified by $\theta \rightarrow T\theta$.

III. *Ω-restriction invariance.* Suppose that F is the family of densities $f(x \mid \theta)$, $\theta \varepsilon \Omega$, and F^* is the family $f(x \mid \theta)$, $\theta \varepsilon \Omega^*$, where Ω^* is an open subset of Ω. We say an inversion g is *Ω-restriction invariant* if for each such F, F^*, $g_{F^*}(\theta \mid x) \propto g_F(\theta \mid x)$, $\theta \varepsilon \Omega^*$.

Essentially this means that the discrimination among θ in Ω^* provided by x is not affected by what values of θ outside Ω^* are possible.

IV. *Sufficiency.* Let $f(x \mid \theta)$, $x \varepsilon S$, $\theta \varepsilon \Omega$ and $f^*(x \mid \theta)$, $x \varepsilon S^*$, $\theta \varepsilon \Omega$ be two families F, F^* of densities; let T be a differentiable but possibly non 1-1 transformation from S to S^*, and $X^* = TX$ be such that the conditional distribution of X given X^* is independent of θ. We say X^* is sufficient for θ; once we know X^* it is irrelevant to θ which particular X gave rise to this value of X^*, since these values of X are distributed, conditionally on X^*, independently of θ.

An inversion g is *sufficiency invariant* if $g_{F^*}(\theta \mid Tx) \propto g_F(\theta \mid x)$. This ensures that the observation $X^* = Tx$ has exactly the same relevance to θ as $X = x$.

V. *Direct product.* Suppose X and Y are independent random variables with densities $f_1(x \mid \theta)$ and $f_2(y \mid \phi)$, respectively, so that the product variable $Z = (X, Y)$ has density $f(x, y \mid \theta, \phi) = f_1(x \mid \theta)f_2(y \mid \phi)$.

Let F_1 be the family of densities $f_1(x \mid \theta)$, $\theta \varepsilon \Omega_1$, let F_2 be the family of densities $f_2(y \mid \phi)$, $\phi \varepsilon \Omega_2$, and let $F = F_1 \times F_2$ be the family of densities $f(x, y \mid \theta, \phi)$, $\theta \varepsilon \Omega_1$, $\phi \varepsilon \Omega_2$.

An inversion g is *direct product invariant* if $g_F(\theta, \phi \mid x, y) \propto g_{F_1}(\theta \mid x) g_{F_2}(\phi \mid y)$.

VI. *Repeated product.* Suppose that $X_1, X_2, \cdots, X_m$ are m independent random variables with the same distribution as X, so that the product random variable $X^* = (X_1, X_2, \cdots, X_m)$ has density $f^*(x_1, x_2, \cdots, x_m \mid \theta) = \prod f(x_i \mid \theta)$.

Let F be the family $f(x \mid \theta)$, $\theta \varepsilon \Omega$ and let F^* be the family $f^*(x_1, \cdots, x_m \mid \theta)$, $\theta \varepsilon \Omega$.

An inversion g is *repetition invariant* if

$$g_{F^*}(\theta \mid x_1, x_2, \cdots, x_m) \propto f(x_2, \cdots, x_m \mid \theta) g_F(\theta \mid x_1),$$

i.e., if g obeys the Bayesian product formula.

Whereas the previous desiderata do not require that g have the form $g_F(\theta \mid x) \propto f(x \mid \theta) h_F(\theta)$, and can reasonably be applied to non-Bayesian inverting methods (e.g. the fiducial method), the repetition invariance property implies that g be determined by $g_F(\theta \mid x) \propto f(x \mid \theta) h_F(\theta)$ for some h (and is then equivalent to requiring that $h_F(\theta) \propto h_{F^*}(\theta)$ in the above terminology).

4. Relatively invariant prior densities. Suppose that h is the kernel of an inversion which is S-labelling, Ω-labelling, and Ω-restriction invariant. Let F be a family of densities $f(x \mid \theta)$, $\theta \varepsilon \Omega$, let Ω_1, Ω_2 be open subsets of Ω and let z be a 1-1 differentiable transformation of S onto S and of Ω_1 onto Ω_2 such that

$$f(zx \mid z\theta)(dzx/dx) = f(x \mid \theta) \qquad \text{for all } x \varepsilon S,\ \theta \varepsilon \Omega_1.$$

Then Properties I, II, III, imply that for some c,

$$h_F(z\theta)(dz\theta/d\theta) = c h_F(\theta) \qquad \text{for all } \theta \varepsilon \Omega_1.$$

Any density which satisfies this relation for all z is a *relatively invariant prior density*. In general there are not enough transformations restraining h_F to determine it uniquely.

The simplest example is the case of symmetric location where F consists of densities $f(x \mid \theta) = k[(x - \theta)^2]$, $-\infty < x < \infty$, $-\infty < \theta < \infty$.

Here transformations $z_c^+ x = x + c$, $z_c^+ \theta = \theta + c$, $z_c^- x = -x + c$, $z_c^- \theta = -\theta + c$, leave F invariant and so the prior density $h_F(\theta)$ satisfies $h_F(\theta + c) = K_c^+ h_F(\theta)$, $h_F(-\theta + c) = K_c^- h_F(\theta)$. In this case θ is uniformly distributed.

Further examples of relatively invariant densities are given in Table 1.

5. Left and right invariant densities. Let F be the family of densities $f(x \mid \theta)$, $x \varepsilon S$, $\theta \varepsilon \Omega$. Consider the set Z of differentiable 1-1 transformations z of S onto S and Ω onto Ω such that

$$f(zx \mid z\theta)(dzx/dx) = f(x \mid \theta).$$

The set Z forms a group under function product. We can say that for $z \varepsilon Z$, zx is to $z\theta$ as x is to θ; therefore $z\theta$ is to zx as θ is to x. This argument has been used explicitly in estimation and testing theory by various authors since 1945 (see Lehmann [5]).

Barnard [1] appears to have been the first to use such an argument in determining prior densities on ignorance. Suppose that Z is simply transitive, i.e., for each θ, $\theta^* \varepsilon \Omega$ there is a unique $z \varepsilon Z$ such that $z\theta = \theta^*$; this means that if we choose some $\theta_0 \varepsilon \Omega$, there is a 1-1 mapping from Z onto Ω given by $z \rightarrow z\theta_0$, and there is a unique group on Ω, with identity θ_0, isomorphic to Z under the mapping.

Barnard suggests that in this case an appropriate prior density on Ω would be the left Haar measure with respect to the group; in our notation he would require that for each z leaving F invariant,

$$h_F(z\theta)(dz\theta/d\theta) = h_F(\theta) \qquad \text{for all} \quad \theta \varepsilon \Omega.$$

If $\theta\theta'$ denotes the product of θ and θ' in the group induced on Ω by Z he would require

$$h_F(\theta'\theta)(d\theta'\theta/d\theta) = h_F(\theta) \qquad \text{for all} \quad \theta', \theta \varepsilon \Omega.$$

Fraser [2], [3], has shown in this case where the group Z is simply transitive, that the fiducial method, with invariant pivotal function, is equivalent to taking a prior density on Ω to be right Haar measure with respect to the induced group and using the Bayesian product formula to obtain the fiducial distribution of θ given x. (In general the fiducial method is not Bayesian.) Specifically he would require

$$h_F(\theta\theta')(d\theta\theta'/d\theta) = h_F(\theta) \qquad \theta, \theta' \varepsilon \Omega.$$

In general these *left and right invariant prior densities* will be different members of the family of invariant prior densities, satisfying, for each z leaving F invariant, for some constant c,

$$h_F(z\theta)(dz\theta/d\theta) = ch_F(\theta) \qquad \text{all} \quad \theta \varepsilon \Omega.$$

6. Locally invariant prior distributions. For general families F there are no transformations z leaving F invariant; even if some transformations exist they rarely determine the prior distribution uniquely. The next three paragraphs introduce a method of assigning prior distributions to families of densities $f(x \mid \theta)$, $\theta \varepsilon \Omega$, Ω an open subset of R, the real line, such that

$$(\partial^r/\partial\theta^r) \log f(x \mid \theta), \quad r \leqq 2, \quad \text{exist for all} \quad x \varepsilon S, \theta \varepsilon \Omega$$

and have finite second moments.

The Ω-restriction property ensures that the value of $h_F(\theta)$ in the neighborhood of θ_0 depends only on the values of $f(x \mid \theta)$ in the same neighborhood of θ_0. $h_F(\theta)$ is determined only up to a constant multiplier so it is appropriate to find the value of $(\partial/\partial\theta) \log h(\theta)$ at $\theta = \theta_0$. This value should depend only on the derivatives $(\partial/\partial\theta^r) \log f(x \mid \theta)$ for $r = 0, 1, 2, \cdots$.

In fact if we could find a transformation T on a neighborhood of θ_0 to another neighborhood of θ_0 such that $T\theta_0 = \theta_0$, and from S onto S, such that

$$f(Tx \mid T\theta)(dTx/dx) = f(x \mid \theta), \qquad x \varepsilon S, \theta \text{ near } \theta_0,$$

we would have

$$h(T\theta)(dT\theta/d\theta) = ch(\theta), \qquad \theta \text{ near } \theta_0,$$

and so at θ_0, $(\partial/\partial\theta) \log h(\theta)$ is given by

$$\left[\frac{\partial T\theta}{\partial\theta}\right]_0 \left[\frac{\partial}{\partial\theta} \log h(\theta)\right]_0 + \frac{\partial}{\partial\theta} \log \frac{dT\theta}{d\theta} = \left[\frac{\partial}{\partial\theta} \log h(\theta)\right]_0. \tag{1}$$

(We are here using $(\partial/\partial\theta)$ as a symbol for ordinary differentiation; $(dT\theta/d\theta)$ means the Jacobian of the transformation $\theta \rightarrow T\theta$; in this case $(dT\theta/d\theta) = |\partial T\theta/\partial\theta|$.)

If we could find a transformation $x \rightarrow Tx$ so that the first three derivatives at θ_0, including the 0th, of the equation

$$\begin{aligned} f(Tx \mid T\theta)(dTx/dx) &= f(x \mid \theta), \quad x \varepsilon S, \theta \text{ near } \theta_0, \text{ namely,} \\ f(Tx \mid \theta_0)(dTx/dx) &= f(x \mid \theta_0), \end{aligned} \tag{2}$$

$$[\partial T\theta/\partial\theta]_0[(\partial/\partial\theta) \log f(Tx \mid \theta)]_0 = [(\partial/\partial\theta) \log f(x \mid \theta)]_0, \tag{3}$$

$$\begin{aligned} [\partial^2 T\theta/\partial\theta^2]_0[(\partial/\partial\theta) \log f(Tx \mid \theta)]_0 &+ [\partial T\theta/\partial\theta]_0^2[(\partial^2/\partial\theta^2) \log f(Tx \mid \theta)]_0 \\ &= [(\partial^2/\partial\theta^2) \log f(x \mid \theta)]_0, \end{aligned} \tag{4}$$

were satisfied for some $[\partial T\theta/\partial\theta]_0$ and $[\partial^2 T\theta/\partial\theta^2]_0$, we could use the values of $[\partial T\theta/\partial\theta]_0$ and $[\partial^2 T\theta/\partial\theta^2]_0$ determined by the equations, to determine $(\partial/\partial\theta) \log h(\theta)$ at $\theta = \theta_0$, using (1).

It now seems plausible to define a *locally invariant prior density* h to be determined at $\theta = \theta_0$ by the solution of (1); namely,

$$(\partial/\partial\theta) \log h(\theta) = T_2/T_1(1 - T_1)$$

if T_1 and T_2 are such that there exists a transformation $x \rightarrow Tx$ with

$$\begin{aligned} [f(Tx \mid \theta)]_0(dTx/dx) &= [f(x \mid \theta)]_0 \\ T_1[(\partial/\partial\theta) \log f(Tx \mid \theta)]_0 &= [(\partial/\partial\theta) \log f(x \mid \theta)]_0 \end{aligned} \tag{5}$$

$$\begin{aligned} T_2[(\partial/\partial\theta) \log f(Tx \mid \theta)]_0 &+ T_1^2[(\partial^2/\partial\theta^2) \log f(Tx \mid \theta)]_0 \\ &= [(\partial^2/\partial\theta^2) \log f(Tx \mid \theta)]_0. \end{aligned} \tag{6}$$

Essentially we are saying that $(\partial/\partial\theta) \log h(\theta)$ depends only on properties of $(\partial/\partial\theta) \log f(x \mid \theta)$ and $(\partial^2/\partial\theta^2) \log f(x \mid \theta)$.

7. Asymptotically locally invariant prior densities. Even if we assume that $(\partial/\partial\theta) \log h(\theta)$ depends only on $(\partial/\partial\theta) \log f(x \mid \theta)$ and $(\partial^2/\partial\theta^2) \log f(x \mid \theta)$, in general we cannot expect to find T_1 and T_2 such that there is a transformation $x \rightarrow Tx$ satisfying (5) and (6).

However the repeatability condition says that the prior density should be the same for the random variable $X = (X_1, X_2, \cdots, X_n)$ with family of densities $\prod f(x_i \mid \theta)$, $\theta \varepsilon \Omega$ as for the random variable X_i with density $f(x_i \mid \theta)$. As $n \to \infty$, the asymptotic distribution of the variables $(\partial/\partial\theta) \log f(x \mid \theta)$, $(\partial^2/\partial\theta^2) \log f(x \mid \theta)$ is determined up to $O(n^{-\frac{1}{2}})$ by the first and second moments of these variables.

Accordingly let us require Equations (5) and (6) to be satisfied up to first and second moments of the variables $(\partial/\partial\theta) \log f(x \mid \theta)$, $(\partial^2/\partial\theta^2) \log f(x \mid \theta)$.

Then we have, writing $f_1 = [(\partial/\partial\theta) \log f(x \mid \theta)]_0$ and $f_2 = [(\partial^2/\partial\theta^2) \log f(x \mid \theta)]_0$,

$$T_1 E(f_1) = E(f_1) \tag{7}$$

$$T_2 E(f_1) + T_1^2 E(f_2) = E(f_2) \tag{8}$$

$$T_1^2 E(f_1^2) = E(f_1^2) \tag{9}$$

$$T_1 T_2 E(f_1^2) + T_1^3 E(f_1 f_2) = E(f_1 f_2) \tag{10}$$

$$T_2^2 E(f_1^2) + 2T_2 T_1^2 E(f_1 f_2) + T_1^4 E(f_2^2) = E(f_2^2). \tag{11}$$

These equations lead to the unique density $(\partial/\partial\theta) \log h(\theta) = -E(f_1 f_2)/E(f_2)$ provided that $E(f_1) = 0$. We call this density *asymptotically locally invariant* (ALI).

8. ALI densities for general parameter spaces. Analogous considerations in the case Ω is an open subset of R^k lead to the following general definition of an ALI density.

DEFINITION. Suppose $f(x \mid \theta)$, $\theta \varepsilon \Omega$, an open subset of R^k, is a family of probability densities, and that at $\theta = \theta_0$ the variables $(\partial/\partial\theta_i) \log f(x \mid \theta)$, $i = 1, \cdot k$, and $(\partial/\partial\theta_i)(\partial/\partial\theta_j) \log f(x \mid \theta)$, $i = 1, \cdot k, j = 1, \cdot k$, have finite first and second moments with

$$E[(\partial/\partial\theta_i) \log f(x \mid \theta)] = 0, i = 1, \cdot k,$$

$$E[(\partial/\partial\theta_i) \log f(x \mid \theta)(\partial/\partial\theta_j) \log f(x \mid \theta)] + E[(\partial^2/\partial\theta_i\partial\theta_j) \log f(x \mid \theta)] = 0,$$

$i = 1, \cdot k, j = 1, \cdot k$. Let g_{ij} be the i, jth element of the inverse of the matrix with i, jth element $E[(\partial/\partial\theta_i)(\partial/\partial\theta_j) \log f(x \mid \theta)]$. Then the ALI (*asymptotically locally invariant*) prior density h is defined at $\theta = \theta_0$ by

$$\frac{\partial}{\partial\theta_p} \log h(\theta) = -\sum_i \sum_j E\left(\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_p} \log f \frac{\partial}{\partial\theta_j} \log f\right) g_{ij},$$

if a solution to these equations exists.

Routine algebra shows that an inversion kernel which assigns to each family of probability densities an ALI prior density generates an inversion which has Properties I, II, III, IV, V, VI.

ALI prior densities have a particularly simple form for the case when $f(x \mid \theta)$ is a member of an exponential family. If

$$f(x \mid \theta) = \exp\left(\sum_{i=1}^{k} l_i(\theta)b_i(x) + L(\theta) + B(x)\right),$$

$\theta \varepsilon \Omega$, an open subset of R^k, is such that $\theta \to (l_1(\theta), l_2(\theta), \cdots, l_k(\theta))$ is a differentiable 1-1 transformation of Ω onto $l(\Omega)$ say, and if the family $f(x \mid \theta)$, $\theta \varepsilon \Omega$ satisfies for all $\theta \varepsilon \Omega$ the regularity conditions detailed in the above definition, the ALI prior density is determined by requiring $(l_1(\theta), \cdots, l_k(\theta))$ to be uniformly distributed over $l(\Omega)$. This simplifies the determination of the prior density in many cases. This prior density for the exponential family was first suggested by Huzurbazar (Jeffreys [5], p. 189).

Examples of ALI densities are given in Table 1.

9. Jeffreys' Prior Density. Jeffreys [4], 1946, first suggested a number of techniques for basing prior distributions on ignorance on properties of the family of densities $f(x \mid \theta)$, $\theta \varepsilon \Omega$. (Perks [7] independently suggested a technique for obtaining prior densities when the parameter space is 1-dimensional, which is equivalent to Jeffreys' principal technique.)

If we let $f_{ij} = E((\partial/\partial\theta_i)(\partial/\partial\theta_j) \log f)$, then Jeffreys' main suggestion was for a priority density $h(\theta) = |f_{ij}|^{\frac{1}{2}}$. This assignment of prior densities generates an inversion which satisfies Properties I, II, III, IV, V, VI. In order to point up the difference between Jeffreys' density and ALI, let us consider the case of a family of densities F such that there exists a 1-1 differentiable transformation z of S onto S, $x \to zx$ and of Ω onto Ω, $\theta \to z\theta$, such that

$$f(zx \mid z\theta)(dzx/dx) = f(x \mid \theta) \qquad \text{for} \quad x \varepsilon S,\ \theta \varepsilon \Omega.$$

Then Jeffreys' density satisfies $h(z\theta)(dz\theta/d\theta) = h(\theta)$ whereas an ALI density satisfies $h(z\theta)(dz\theta/d\theta) = ch(\theta)$ for some constant c. The difference therefore is between left invariance and relative invariance. In fact if we pursue the heuristic arguments of Sections 6 and 7 requiring that the prior density be left invariant, we will obtain that the existence of a transformation $x \to Tx$ such that

$$f(Tx \mid \theta_0)\,\frac{dTx}{dx} = f(x \mid \theta_0), \qquad \frac{\partial T\theta}{\partial\theta}\,\frac{\partial}{\partial T\theta} \log f(Tx \mid T\theta) = \frac{\partial}{\partial\theta} \log f(x \mid \theta),$$

suggests $h(T\theta)(dT\theta/d\theta) = h(\theta)$, where θ and $T\theta$ are any two values of θ and $\partial T\theta/\partial\theta$ is any constant.

Asymptotically this reduces to

$$(dT\theta/d\theta)^2 E[((\partial/\partial T\theta) \log f)^2] = E[((\partial/\partial\theta) \log f)^2]$$

implies

$$h(T\theta)(dT\theta/d\theta) = h(\theta), \quad \text{or} \quad h(\theta) = \{E[((\partial/\partial\theta) \log f)^2]\}^{\frac{1}{2}}$$

which is Jeffreys' form. An easy generalization extends the argument to general parameter spaces.

If $J(\theta)$ denotes Jeffreys' prior density and $H(\theta)$ the ALI prior density,

$J(\theta)^{\alpha}H(\theta)^{\beta}$, where $\alpha + \beta = 1$, is an assignment of prior densities satisfying all the usual properties.

10. Suggestions for use. If we use the family $f(x \mid \theta)$ to determine the prior density on ignorance, $h(\theta)$, we must restrict the circumstances under which we can use the usual Bayesian product law. An attractive property of the Bayesian technique is that the posterior distribution of θ given the observations $x_1, \cdots, x_n$ summarizes the information about θ given by the observations; if we make some further observations $y_1, y_2, \cdots, y_m$ we can calculate the posterior distribution of θ given $x_1, x_2, \cdots, x_n, y_1, \cdots, y_m$ by using Bayes' product law

$$g(\theta \mid x, y) \propto f(y \mid \theta)g(\theta \mid x),$$

with prior distribution equal to the posterior distribution of θ given $x_1, \cdots, x_n$. If we base the prior density on ignorance on the form of $f(x \mid \theta)$ we can no longer consistently use the product rule in this way unless the x and y have the same family of probability densities. To illustrate suppose we obtain r successes in n trials from a binomial distribution with parameter θ; the ALI prior density is $1/(\theta(1 - \theta))$. We then sample from the binomial distribution until r' successes are obtained; suppose n' trials are necessary; the ALI prior is $1/(1 - \theta)$. If we find the posterior distribution for θ given r, n first we get $g(\theta \mid r, n) \propto \theta^{r-1}(1 - \theta)^{n-r-1}$ and using the product formula $g(\theta \mid r, n, r', n') \propto \theta^{r+r'-1}(1 - \theta)^{n+n'-r-r'-1}$. If we find the posterior distribution for θ given r', n' first we get $g(\theta \mid r', n') \propto \theta^{r}(1 - \theta)^{n-r-1}$ and so $g(\theta \mid r, n, r', n') \propto \theta^{r+r'}(1 - \theta)^{n+n'-r-r'-1}$. In particular the posterior distribution obtained using these techniques may depend on the "sampling rule".

The referee has pointed out an apparently undesirable property of the rule ALI. Suppose $X_1, X_2, \cdots, X_n$ are observations from $N(0, \sigma^2)$; then

$$f(x \mid \sigma) = ((2\pi)^{-\frac{1}{2}})^n \exp\left(-\tfrac{1}{2} \sum x^2/\sigma^2\right)$$

leads to an ALI prior $h(\sigma) = 1/\sigma^3$, and the posterior distribution of σ given $x_1, \cdots, x_n$ is such that $\sum x_i^2/\sigma^2$ is χ^2 with $(n + 2)$ degrees of freedom.

If $X_1, \cdots, X_n$ are observations from $N(\mu, \sigma^2)$, then

$$f(x \mid \mu, \sigma) = ((2\pi)^{-\frac{1}{2}})^n \exp\{-\tfrac{1}{2} \sum [(x - \mu)/\sigma]^2\}$$

leads to an ALI prior $h(\sigma, \mu) = 1/\sigma^5$, and the posterior distribution of σ given $x_1, \cdots, x_n$ is such that $\sum (x - \bar{x})^2/\sigma^2$ is χ^2 with $(n + 3)$ degrees of freedom (where $n\bar{x} = \sum x$). This is contrary to the ordinary association of degrees of freedom with information; for we lose a degree of freedom if we have information that $\mu = 0$.

Jeffreys' prior yields a posterior distribution of σ given $x_1, \cdots, x_n$ such that $\sum x^2/\sigma^2$ is χ^2 with n degrees of freedom in the case $N(0, \sigma^2)$, and $\sum (x - \bar{x})^2/\sigma^2$ is χ^2 with n degrees of freedom in the case $N(\mu, \sigma^2)$; Jeffreys has himself expressed dissatisfaction with his rule in this case.

TABLE 1

Examples of the various invariant prior densities

Family of Densities			Type of invariant prior density				
Form	Sample Space	Parameter Space	Relative	Left	Right	Jeffreys'	ALI
$(2\pi)^{-\frac{1}{2}} \exp [-\frac{1}{2}(x - \theta)^2]$	$-\infty < x < \infty$	$-\infty < \theta < \infty$	1	1	1	1	1
$\theta^{-1}(2\pi)^{-\frac{1}{2}} \exp (-\frac{1}{2}x^2/\theta^2)$	$-\infty < x < \infty$	$0 < \theta < \infty$	θ^k	θ^{-1}	θ^{-1}	θ^{-1}	θ^{-3}
$\theta_2^{-1}(2\pi)^{-\frac{1}{2}} \exp [-\frac{1}{2}(x - \theta)^2/\theta_2^2]$	$-\infty < x < \infty$	$-\infty < \theta < \infty$, $0 < \theta_2 < \infty$	θ_2^k	θ_2^{-2}	θ_2^{-1}	θ_2^{-2}	θ_2^{-5}
$\lvert\theta\rvert^{\frac{1}{2}}(2\pi)^{-\frac{1}{2}n} \exp (-\frac{1}{2}x'\theta x)$	$x \in R^n$	θ positive definite $n \times n$ matrix	$\lvert\theta\rvert^k$	$\lvert\theta\rvert^{-1}$	*	$\lvert\theta\rvert^{-1}$	1
$(2\pi)^{-\frac{1}{2}n} \exp [-\frac{1}{2}(x - K\theta)'(x - K\theta)]$ K, $n \times k$ matrix of rank k	$x \in R^n$	$\theta \in R^k$	1	1	1	1	1
$x^{\theta-1}e^{-x}/\Gamma(\theta)$	$0 < x < \infty$	$0 < \theta < \infty$	*	*	*	$[(d^2/d\theta^2) \log \Gamma(\theta)]^{\frac{1}{2}}$	1
$\theta^x e^{-\theta}/x!$	x non-negative integer	$0 < \theta < \infty$	*	*	*	$\theta^{-\frac{1}{2}}$	θ^{-1}
$\theta_1^{x_1}\theta_2^{x_2} \cdots \theta_r^{x_r} n!/x_1 \cdots x_r!$	x_i integers $x_i \geqq 0$, $\sum x_i = 1$	$\theta_i > 0$, $\sum \theta_i = 1$	*	*	*	$(\theta_1\theta_2 \cdots \theta_r)^{-\frac{1}{2}}$	$(\theta_1\theta_2 \cdots \theta_r)^{-1}$
$\theta^x(1 - \theta)^{n-x} n!/x!(n - x)!$	x integer, $0 \leqq x \leqq n$	$0 < \theta < 1$	*	*	*	$[\theta(1 - \theta)]^{-\frac{1}{2}}$	$[\theta(1 - \theta)]^{-1}$
$\theta^r(1 - \theta)^{x-r}(x - 1)!/(r - 1)!(x - r)!$	x integer, $x \geqq r$	$0 < \theta < 1$	*	*	*	$(1 - \theta)^{-\frac{1}{2}}\theta^{-1}$	$(1 - \theta)^{-1}$

* Means either that the method is not defined for the family of densities considered, or that it does not determine a prior density for the family.

The composite prior density J^2H^{-1} would yield $\sum x^2/\sigma^2$ is χ^2 with $(n - 2)$ degrees of freedom in the case $N(0, \sigma^2)$ and $\sum (x - \bar{x})^2/\sigma^2$ is χ^2 with $(n - 3)$ degrees of freedom in case $N(\mu, \sigma^2)$; this is simply a makeshift device to deal with this particular anomaly.

In conclusion, there seem compelling reasons for requiring a prior density to be relatively invariant. This will frequently not restrain it much. In selecting a specific prior density there are less compelling heuristic reasons to use ALI; any member of the family $J^\alpha H^\beta$, $\alpha + \beta = 1$, has Properties I, II, $\cdots$, VI and is therefore worthwhile considering.

11. Acknowledgment. The author is indebted to Professors F. J. Anscombe and J. W. Tukey of Princeton University for advice and encouragement during the preparation of the material presented in this paper, and to the referee for his valuable suggestions and comments.

REFERENCES

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

[2] FRASER, D. A. S. (1961). On fiducial inference. *Ann. Math. Statist.* **32** 661–676.

[3] FRASER, D. A. S. (1961). The fiducial method and invariance. *Biometrika* **48** 261–280.

[4] JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser A* **186** 453–461.

[5] JEFFREYS, H. (1960). *Theory of Probability*, (3rd ed.). Oxford Univ. Press.

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

[7] PERKS, F. J. A. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries* **73** 285–334.