Biometrika (1966), **53**, 1 *and* 2, *p.* 1
Printed in Great Britain

Structural probability and a generalization*

By D. A. S. FRASER
University of Toronto

Summary

Structural probability, a reformulation of fiducial probability for transformation models, is discussed in terms of an error variable. A consistency condition is established concerning conditional distributions on the parameter space; this supplements the consistency under Bayesian manipulations found in Fraser (1961). An extension of structural probability for real-parameter models is developed; it provides an alternative to the local analysis in Fraser (1964*b*).

1. Introduction

Fiducial probability has been reformulated for location and transformation models (Fraser, 1961) and compared with the prescriptions in Fisher's papers (Fraser, 1963*b*). The transformation formulation leads to a frequency interpretation and to a variety of consistency conditions; the term *structural probability* will be used to distinguish it from Fisher's formulation.

Fiducial probability was introduced by Fisher in 1930 and developed along with other inference methods through many of his papers. Fisher's work in inference seems to be the main basis and stimulus for the present attention to aspects of inference, such as likelihood, conditionality, significance testing, and seems to have led thereby to the present substantial alternatives to decision theory. Dempster (1964) is somewhat alone in belittling this large contribution. His criticisms seem most to indicate disillusionment that Fisher's contributions are not organically whole and logically consistent in codified form. This is certainly ignoring the magnitude of the actual contributions, but perhaps more dangerously it is ignoring possible developments from things but lightly touched by Fisher.

In commenting on fiducial probability for transformation models Dempster remarks '... it simply transforms the problem of choosing pivoted variables into a slightly narrower problem of choosing a group, and the latter problem seems to me unintuitive and far removed from the central issue'. The basis for this opinion would seem to be against the *standard statistical model*: a variable, a parameter, a probability density function, and nothing more. But in many applications there is more and it should appear as additional structure in the statistical model. Some examples will be considered in this section. Against the augmented model Dempster's remark is inappropriate.

The simple measurement model in ordinary form involves a variable x, a parameter θ and a probability density function $f(x-\theta)$ with f specified. In analyses based on this, there is emphasis on x as a *variable* and on θ as *fixed*. The location group as used in Fraser (1961) provides a simple means of describing the fixed shape of the distribution with differing values for θ.

* Prepared at the University of Copenhagen 1964; revised at the Mathematics Research Center, United States Army and at the Department of Statistics, University of Wisconsin.

The simple measurement model can be presented alternatively in the form

$$x = \theta + e$$

where e, an error variable, has a fixed distribution given by $f(\cdot)$. In an analysis involving a particular instance, emphasis can be placed on x as *known* and θ as *unknown*. Particular values for x and θ would depend on the conventional origin on the measurement scale and can be viewed separately from the error variable with its known distribution. A value for e gives the position of x with respect to θ and its negative gives the position of θ with respect to x. The ordinary form for this model has gone too far in making distinctions—a different variable for a different θ—and needs the group to partially recover the essential variable by a symmetry argument.

Consider the measurement model in the alternative form

$$x = \theta + e.$$

In a particular instance there is a realized value e^r for the error variable. Probability statements concerning the unobserved e^r can be made in exactly the way a dealer at poker might calculate probabilities after dealing cards but before observing any. The probability element

$$f(e)\,de$$

describes the distribution appropriate to e^r. With x observed and *known* and θ *unknown*, the distribution of $-e$ which describes θ position with respect to x provides the distribution of possible values for θ

$$f(x-\theta)\,d\theta,$$

the *structural distribution* for θ based on an observed x.

The referee of the original version of this paper commented 'I cannot see how, within the ordinary meaning of probability, a distribution for θ can be obtained. An extra principle is needed in order to transfer the distribution from e to θ. This principle is not contained within the ordinary probability calculus. My feeling is that the principle should be stated explicitly'.

For someone committed to the ordinary statistical model, an extra principle is seemingly needed. But its introduction can only be to compensate for an inadequacy in the ordinary model. For the alternative form of the measurement model, an extra principle is not needed: *x is known, θ is unknown, the distribution of θ position with respect to x is known, and the origin of the scale of measurement is conventional.*

With multiple observations the measurement model takes the form

$$\begin{aligned} x_1 &= \theta + e_1 \\ &\vdots \\ x_n &= \theta + e_n, \end{aligned}$$

where $(e_1, \ldots, e_n)$ is a vector sample of error variables from the distribution $f(e)$. In a particular instance there is a realized sample $(e_1^r, \ldots, e_n^r)$ from the error distribution. Some aspects of the realized errors *are* observable

$$\begin{aligned} e_2^r - e_1^r &= x_2 - x_1 \\ &\vdots \\ e_n^r - e_1^r &= x_n - x_1. \end{aligned}$$

One aspect, however, is *not* observable; it describes the location of the realized errors and can be described by, say, e_1^r

$$x_1 = \theta + e_1^r.$$

Probability statements concerning the unobserved e_1^r can be made in exactly the way a dealer at poker might calculate probabilities after dealing the cards and after observing his own hand: the dealer would condition on the cards observed. Correspondingly, the statistician should condition on the observed components of the error sample; the probability element for e_1 given $e_2^r - e_1^r, \ldots, e_n^r - e_1^r$ is

$$g(e_1 \mid e_2^r - e_1^r, \ldots, e_n^r - e_1^r)\, de_1 = \frac{f(e_1) f(e_1 + e_2^r - e_1^r) \ldots f(e_1 + e_n^r - e_1^r)}{h(e_2^r - e_1^r, \ldots, e_n^r - e_1^r)}\, de_1,$$

where $h(y_2, \ldots, y_n)$ is the marginal density for the sample differences $e_2 - e_1, \ldots, e_n - e_1$. With x_1 known and θ unknown, the conditional distribution of $-e_1$, which describes θ position with respect to x provides the distribution of possible values for θ

$$g(x_1 - \theta \mid x_2 - x_1, \ldots, x_n - x_1)\, d\theta = c(x_1, \ldots, x_n) \prod_{i=1}^{n} f(x_i - \theta)\, d\theta$$

the *structural distribution* for θ based on an observed sample $(x_1, \ldots, x_n)$.

For the general case consider an error variable e on a space X. And suppose that an observable x is obtained by a transformation $[\theta]$ applied to the variable e

$$x = [\theta]\, e.$$

Suppose that the transformations $[\theta]$ are indexed by a parameter θ with values in a parameter space Ω and that the transformations $[\theta]$ are precisely the transformations of a group $G = \{g\}$ that is *unitary* in its application to the space X: if $gx = hx$ then $g = h$ (at most one transformation carrying any point into any other point).

The elements of G and Ω are in one-one correspondence:

$$[\theta] \longleftrightarrow \theta.$$

Any element of G can produce a transformation on G by left multiplication; correspondingly there is a transformation on Ω. Let the same group element be used to designate this isomorphic transformation. This, then, permits the representation

$$\theta = [\theta]\, \theta_0,$$

where the identity element is in correspondence with θ_0: $e \longleftrightarrow \theta_0$.

To avoid degeneracy suppose that $[\theta']e$ and $[\theta'']e$ have different distributions whenever $\theta' \neq \theta''$.

The measurement model with unknown scaling provides a simple example of this general model

$$\begin{aligned} x_1 &= \mu + \sigma e_1 \\ &\vdots \\ x_n &= \mu + \sigma e_n, \end{aligned}$$

where $(e_1, \ldots, e_n)$ is a vector sample from a distribution $f(e)$. This can be represented in the form

$$\mathbf{x} = [\mu, \sigma]\, \mathbf{e},$$

where a transformation $[a, c]$ is a location-scale transformation

$$[a, c]\, (x_1, \ldots, x_n) = (a + cx_1, \ldots, a + cx_n)$$

with $-\infty < a < \infty$, $0 < c < \infty$. In this example a latent error distribution is relocated and rescaled; in a typical application this would reflect the conventional nature of the origin and unit of measurement.

For a model described by a density function it is natural to use a measure that has invariance under the group of transformations. Let M be an invariant measure on X: $M(A) = M(gA)$ for all g in G and all A contained in X; let μ be the left invariant measure on G: $\mu(H) = \mu(gH)$; and let Δ be the modular function satisfying

$$d\mu(g) = \Delta(g)\, d\mu(g^{-1}),$$
$$d\mu(gh) = \Delta(h)\, d\mu(g),$$

where g is the measure variable.

Suppose now that e has a probability density function f with respect to the invariant measure M. The general model then has the form

$$x = [\theta]e$$

where the error variable e has element $f(e)\, dM(e)$.

Under transformations in G applied to X a point x is carried into an orbit

$$S_x = \{gx \mid g \in G\}.$$

Suppose that a reference point is chosen on each orbit and let $[x]$ designate the unique transformation in G that carries the reference point on the orbit S_x into the point x. The reference point for the orbit through x can then be designated by

$$[x]^{-1}x.$$

And since there is precisely one reference point on each orbit, the expression $[x]^{-1}x$ can be used to label the orbit through x.

In a particular instance there is a realized value e^r from the error variable. Some aspects of the realized error are observable

$$\begin{aligned}[e^r]^{-1}e^r &= [e^r]^{-1}[\theta]^{-1}[\theta]e^r \\ &= [[\theta]e^r]^{-1}[\theta]e^r \\ &= [x]^{-1}x.\end{aligned}$$

One aspect, however, is not observable; it describes the location of e^r on the orbit $[e^r]^{-1}e^r$ and is given conveniently by $[e^r]$

$$[x] = [\theta][e^r].$$

Probability statements concerning $[e^r]$ are obtained from the distribution of $[e]$ conditional on the observed aspects $[e^r]^{-1}e^r$; this conditional distribution (Fraser, 1963*b*, §4·2) has element

$$k([e^r]^{-1}e^r) f([e]\,.\,[e^r]^{-1}e^r)\, d\mu([e]),$$

where k is a normalizing constant on the orbit $[e^r]^{-1}e^r$. This conditional distribution can be transformed to obtain the distribution of $[e]^{-1}$, which describes θ position with respect to x

$$[\theta] = [x][e]^{-1}.$$

With x known and θ unknown the probability element can be manipulated:

$$\begin{aligned}&k([e^r]^{-1}e^r) f([e][e^r]^{-1}e^r)\, d\mu([e]) \\ &\quad = k([e^r]^{-1}e^r) f([e][e^r]^{-1}e^r)\, \Delta([e])\, d\mu([e]^{-1}) \\ &\quad = k([x]^{-1}x) f([\theta]^{-1}[x][x]^{-1}x)\, \Delta([\theta]^{-1}[x])\, d\mu([x]^{-1}[\theta]) \\ &\quad = k([x]^{-1}x) f([\theta]^{-1}x)\, \Delta([\theta]^{-1}[x])\, d\mu([\theta]);\end{aligned}$$

this is the *structural distribution* for the unknown θ based on a known x.

For the measurement model with unknown scale parameter take as reference point on the orbit

$$S_{\mathbf{x}} = \{(a+cx_1, \ldots, a+cx_n)\},$$

the point having $\bar{x} = 0$ and $s_x = 1$; then

$$[\mathbf{x}] = [\bar{x}, s_x].$$

The observable aspect of the error is

$$[\bar{x}, s_x]^{-1}\mathbf{x} = \left(\frac{x_1-\bar{x}}{s_x}, \ldots, \frac{x_n-\bar{x}}{s_x}\right)$$

$$= \left(\frac{e_1-\bar{e}}{s_e}, \ldots, \frac{e_n-\bar{e}}{s_e}\right)$$

and the unobservable aspect is $(\bar{e}, s_e]$. From Fraser (1963*b*) the conditional distribution of $[\bar{e}, s_e]$ is

$$k\left(\frac{e_i^r-\bar{e}^r}{s_e^r}\right)\Pi f\left([\bar{e}, s_e]\frac{e_i^r-\bar{e}^r}{s_e^r}\right)s_e^n\frac{d\bar{e}\,ds_e}{s_e^2},$$

and the structural distribution for $[\mu, \sigma]$ is

$$k\left(\frac{x_i-\bar{x}}{s_x}\right)\Pi f\left(\frac{x_i-\mu}{\sigma}\right)\left(\frac{s_x}{\sigma}\right)^n\left(\frac{\sigma}{s_x}\right)\frac{d\mu\,d\sigma}{\sigma^2}.$$

2. Consistency: conditional distributions

Some consistency properties of structural probability have been examined in Fraser (1961, 1962): that the structural distribution from one set of variables can be used as a prior distribution for a Bayesian analysis on another set of variables with a result independent of the choice of the first set (provided all variables generate the same transformation group on the parameter space); and that a structural distribution can be combined directly with a prior distribution and yield the same result as a Bayesian analysis. In this section a consistency property for conditioned structural distributions is considered.

Consider first an example. Let $(x_1, \ldots, x_n)$ be a sample from the model $x = \mu + \sigma e$ where e is standard normal. The structural distribution for (μ, σ) can be represented by

$$\mu = \bar{x} - \frac{z}{(n-1)^{-\frac{1}{2}}\chi}n^{-\frac{1}{2}}s,$$

$$\sigma = \frac{s}{(n-1)^{-\frac{1}{2}}\chi},$$

where z and χ are independent and are respectively standard normal and chi on $n-1$ degrees of freedom.

Suppose the information $\sigma = \sigma_0$ becomes available. The joint distribution for (μ, σ) is easily conditioned since $\sigma = \sigma_0$ implies that

$$\chi = (n-1)^{\frac{1}{2}}\frac{s}{\sigma}$$

is known in value; and since z is statistically independent of χ it follows that

$$\mu = \bar{x} - \frac{z\sigma_0}{s} n^{-\frac{1}{2}} s$$

$$= \bar{x} - z n^{-\frac{1}{2}} \sigma_0.$$

This conditioned distribution is exactly the structural distribution that is obtained from the model: $(x_1, \ldots, x_n)$ is a sample from $x = \mu + \sigma_0 e$ where e is standard normal.

Alternatively, suppose that the information $\mu = \mu_0$ becomes available. The joint probability element for (μ, σ)

$$\frac{A_{n-1}}{(2\pi)^{\frac{1}{2}n}} \exp\left\{-\frac{n(\bar{x}-\mu)^2 + (n-1)s^2}{2\sigma^2}\right\} \left((n-1)^{\frac{1}{2}} \frac{s}{\sigma}\right)^n \left(\frac{n}{n-1}\right)^{\frac{1}{2}} \frac{\sigma}{s} \frac{du\, d\sigma}{\sigma^2}$$

can be conditioned according to $\mu = \mu_0$ and yields

$$k \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \left(\frac{S}{\sigma}\right)^n \frac{d\sigma}{\sigma}$$

where $S^2 = n^{-1}\Sigma(x_i - \mu_0)^2$. This is the structural distribution as obtained from the model: $(x_1, \ldots, x_n)$ is a sample from $x = \mu_0 + \sigma e$ where e is standard normal.

For the general case consider the transformation model (X, G, Ω):

$$x = [\theta] e,$$

where e has the element $f(e)\, dM(e)$

on X, where $[\theta]$ takes values in the transformation group G unitary on X, and where $\theta = [\theta]\theta_0$ takes values in Ω.

THEOREM. *If (X, H, Ω_0) is a transformation model with H a subgroup of G and $\Omega_0 = H\theta_0$, then the structural distribution from the submodel is the same as the structural distribution from the full model as conditioned to Ω_0 with respect to partition sets $g\Omega_0$.*

Proof. A structural distribution derives from a conditional distribution on an orbit. For the model (X, G, Ω) it suffices then to suppose that the sample space X consists of just one orbit. Let x_0 be a reference point in the sample space and let the error $e = gx_0$ be expressed in terms of a variable g on the group G. And let θ_0 be a reference point in Ω; for convenience take θ_0 in Ω_0.

The subgroup H generates orbits on the space X. Let $x_0(x)$ be a reference point on the orbit through the point x and let

$$x_0(x) = a_x x_0,$$

where a_x is an element of G; x can then be written

$$x = h_x a_x x_0,$$

where h_x is an element of H. The error variable e can correspondingly be expressed in terms of components

$$e = gx_0 = hax_0,$$

where a has the marginal distribution of the orbital variable and h has a conditional distribution given a. The full model can then be expressed in the form

$$x = [\theta] g x_0 = [\theta] h a x_0.$$

The structural distribution for the full model has the form

$$\theta = [x]g^{-1}\theta_0$$
$$= [x]a^{-1}h^{-1}\theta_0.$$

This is a distribution $(h^{-1}\theta_0)$ on Ω_0 transformed by a random group element $f = [x]a^{-1}$; it is thus expressed in a form appropriate to the partition $f\Omega_0$. The conditional distribution on Ω_0 is then obtained from the condition $a = a_x$ and has the form

$$\theta = h_x h^{-1}\theta_0,$$

where h has the conditional distribution given $a = a_x$.

Consider now the submodel with group H. The orbits are given by a_x and position on an orbit by h_x. The structural distribution then has the form

$$\theta = h_x h^{-1}\theta_0,$$

where h has the conditional error distribution on the orbit given by a_x. The two structural distributions are the same.

3. An extension of structural probability

For models involving a real variable and a real parameter, structural probability as discussed in the preceding sections is available only for the location model $f(x-\theta)$. One extension for stochastically increasing variables is considered in Fraser (1964*b*); it leads to local structural probability and a residual likelihood and is based on local conditional sufficiency (Fraser, 1964*a*). In this section an alternative extension will be considered; it leads to a global structural distribution but with possible non-uniqueness dependent on the choice of initiating variable.

Consider a real variable x with a stochastically increasing distribution:

$$F_\theta(x \mid \theta) = \frac{\partial}{\partial\theta}F(x \mid \theta) < 0.$$

Properties of the distribution for x near x_0 will be used to analyse the parameter space. First, consider the increment x_0, $x_0 + d$; the distribution function increases in value by the amount $F_x(x_0 \mid \theta)\, d$. Next, consider the increment $\theta, \theta + \delta$; the distribution function decreases by an amount $-F_\theta(x_0 \mid \theta)\, \delta$. The distribution function value $F(x_0 \mid \theta)$ will be approximately equal to $F(x_0 + h, \theta + \delta)$ if δ and h are in the ratio given by

$$F_x(x_0 \mid \theta)\, h = -F_\theta(x_0 \mid \theta)\, \delta.$$

A change in x at x_0 can thus be viewed as corresponding to a topological shift on the parameter space with rate at θ given by

$$\frac{1}{h(\theta, x_0)} = -\frac{F_x(x_0 \mid \theta)}{F_\theta(x_0 \mid \theta)};$$

the function h is, in a sense, the density of θ values with respect to such a shift. A new parameter $\tau(\theta)$ can be defined having a rate of shift equal to unity:

$$\frac{d\theta(\tau)}{d\tau} = -\frac{F_x(x_0 \mid \theta)}{F_\theta(x_0 \mid \theta)},$$

$$\tau = \int^\theta -\frac{F_\theta(x_0 \mid \theta)}{F_x(x_0 \mid \theta)}\, d\theta = \int^\theta h(\theta, x)\, d\theta.$$

Let $H(x, \tau) = F(x \mid \theta(\tau))$ be the distribution function in terms of the transformed parameter. The definition of τ then shows that in the neighbourhood of x_0 the distribution has location form with respect to τ.

Consider now structural inference based on the local location form of the distribution. For an observation x let τ be the transformed parameter corresponding to the neighbourhood x. The structural probability element is then

$$|H_\tau(x \mid \tau)|\, d\tau = |F_\theta(x \mid \theta)|\, d\theta;$$

the structural density function for θ has the form

$$\begin{aligned} -F_\theta(x \mid \theta) &= F_x(x \mid \theta)\left(-\frac{F_\theta(x \mid \theta)}{F_x(x \mid \theta)}\right) \\ &= F_x(x \mid \theta)\, h(\theta, x) \end{aligned}$$

and is thus seen to be the likelihood function $F_x(x \mid \theta)$ modulated by the function $h(\theta, x)$.

Consider now a sample $(x_1, \ldots, x_n)$ from the distribution $F(x \mid \theta)$. The structural distribution from x_1 has the density

$$F_x(x_1 \mid \theta)\, h(\theta, x_1).$$

Using this in a Bayesian analysis on the remainder of the sample yields the following *relative* density function* for θ

$$\Pi F_x(x_i \mid \theta)\, h(\theta, x_1)$$

with a normalizing constant, say $c_1(\mathbf{x})$, that can be determined by integration; it is the joint likelihood function modulated by the θ-density function from the first observation. This distribution has a frequency interpretation in terms of the local structure of the distribution at the first observation.

Alternatively, commencing from the observation x_j a structural distribution with relative density

$$\Pi F_x(x_i \mid \theta)\, h(\theta, x_j)$$

is obtained.

The criterion of uniqueness has been applied more severely to fiducial probability than to other areas of inference, perhaps in part because of Fisher's claim of uniqueness. In structural probability the criterion is in large measure satisfied for transformation models. The criterion may, however, be too strong to invoke for more general models. An inference analysis for a stochastically increasing model $F(x \mid \theta)$ might examine each of the structural distributions

$$c_j(\mathbf{x})\, h(\theta, x_j)\, \Pi F_x(x_i \mid \theta)$$

and in some contexts might go further and examine weighted combinations

$$\Sigma l_j c_j(\mathbf{x})\, h(\theta, x_j)\, \Pi F_x(x_i \mid \theta),$$

in particular the symmetric combination

$$\Sigma n^{-1} c_j(\mathbf{x})\, h(\theta, x_j)\, \Pi F_x(x_i \mid \theta).$$

In each of these the likelihood function is present and is modulated by a function based on the θ-densities $h(\theta, x_j)$. As noted by the referee these extended structural distributions will typically violate the likelihood principle; for the author this is not viewed as being adverse to the extension (Fraser, 1963*a*).

* A distribution of this form has been proposed by Roy (1960).

References

DEMPSTER, A. P. (1964). On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Ass.* **59**, 56–66.

FISHER, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.* **26**, 528–38.

FRASER, D. A. S. (1961). The fiducial method and invariance. *Biometrika*, **48**, 261–80.

FRASER, D. A. S. (1962). On the consistency of the fiducial method. *J. R. Statist. Soc.* B, **24**, 425–34.

FRASER, D. A. S. (1963*a*). On the sufficiency and likelihood principles. *J. Amer. Statist. Ass.* **58**, 641–7.

FRASER, D. A. S. (1963*b*). On the definition of fiducial probability. *Bull. Int. Statist. Inst.* **40**, 842–56.

FRASER, D. A. S. (1964*a*). Local conditional sufficiency. *J. R. Statist. Soc.* B, **26**, 52–62.

FRASER, D. A. S. (1964*b*). On local inference and information. *J.R. Statist. Soc.* B, **26**, 253–60.

ROY, A. D. (1960). Some notes on pistimetric inference. *J. R. Statist. Soc.* B, **22**, 338–47.