The Annals of Mathematical Statistics
1972, Vol. 43, No. 3, 777-790

BAYES, LIKELIHOOD, OR STRUCTURAL[1]

By D. A. S. Fraser

University of Hawaii and University of Toronto

0. Summary. The traditional model of statistics is examined in Section 1. The model, as such, distinguishes response values only by their likelihood functions. Recognition of this restricted identification effectively precludes the need for any theory of sufficient statistics.

A probability space that has an attached observer-processor-reporter (OPR) mechanism is examined in Section 2 as a means of assessing the nature of reported information; such information may or may not be observational in character depending on properties of the OPR mechanism.

A variation-response model is a probability space and a class of random variables: the probability space describes the sources of variation in the system under investigation and the class of random variables describes the possible presentations of this variation in the response of the system. Section 3 examines how realized values on the probability space are distinguished or identified by the model; Section 4 considers how distributions on the probability space are identified by response variable data.

In Sections 5, 6, 7 the essentials of three contemporary approaches to inference are presented and each is accompanied by criticisms that proponents of the other methods might make. The key to these criticisms lies primarily in whether hypothetical information is added to or substantiated information is omitted from the assembled information concerning the system under investigation.

In certain contexts the bayesian approach makes an arbitrary but typically consistent choice of input to its analyses; it is not the input suggested by standard invariance analysis. In those cases where a variation-response model is appropriate, the use of this more embracing model presents theoretical support for the bayesian choice (Section 8); but of course with the more embracing model the bayesian premises are not needed to obtain the usual bayesian result.

An example in Section 9 illustrates the theoretical simplicity of classical probabilities for certain unknowns other than realized values on a probability space.

1. Outcome identification by the traditional model. Consider a system $\mathscr{E}$ under investigation: let x be the the response variable, $\mathscr{X}$ be the response space, and $\mathscr{A} = \{A\}$ be the σ-algebra of permissible events A for the response variable. A *traditional model* in its most abstracted form is a class $C = \{P_\theta : \theta \in \Omega\}$ of probability measures on $(\mathscr{X}, \mathscr{A})$.

In developing a model for the system $\mathscr{E}$ the investigator assembles the

Received November 3, 1969; revised November 3, 1971.

[1] Supported in part by Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

information concerning the system $\mathscr{E}$ and concerning related systems as they bear on $\mathscr{E}$; such information effectively presents the possibilities for what could be (as opposed to what could not be) all within a reasonable degree of approximation. A traditional model C is called *relevant* for $\mathscr{E}$ if the distributions in C are effectively the response-variable distributions in the assembled information. A traditional model C is called *adequate* for $\mathscr{E}$ if it is relevant for $\mathscr{E}$ and if the distributions in C are effectively *all* the assembled information. Let θ_0 designate the distribution describing the response. And let x_0 designate the realized response value from a performance of the system. The problem of inference then is to determine from an adequate model C the information about θ_0 provided by x_0 with C.

For an adequate model C consider how the response values in $\mathscr{X}$ are distinguished one from another by the model. A value x is identified in the model C by a set of functions of θ,

$$S_1(x) = \{P_\theta(A) : x \in A\} ,$$

the probabilities for realized events as functions of θ. The set $S_1(x)$ is all that the model presents concerning the value x, and is thus all that model provides for making distinctions among response values. The problem of inference thus becomes *the analysis of* $S_1(x_0)$ *with* C.

In most cases with any regularity the set $S_1(x)$ can be represented by functions referring to limiting neighbourhoods, what might be called germs of $S_1(x)$. For example with a discrete model C_d on the finest σ-algebra $\mathscr{A}$ over $\mathscr{X}$, a value x is identified by a single function of θ,

$$S_2(x) = P_\theta(\{x\}) .$$

This function is all that the model provides for making distinctions among response values. The problem of inference thus becomes *the analysis of* $S_2(x)$ *with* C. As a second example consider a continuous model C_c: suppose the σ-algebra $\mathscr{A}$ is based on the open sets of a metrizable topology and that the derivative of each P_θ relative to some atom-free measure m exists at each point x,

$$\lim_{x \in A; d(A) \to 0} \frac{P_\theta(A) - f(x : \theta)m(A)}{m(A)} = 0$$

where $d(A)$ is the diameter of A relative to some supporting metric. A value x is thus identified in the model C_c by a measure differential $f(x : \theta)m$ with m unspecified in positive value,

$$S(x) = \{f(x : \theta)m : m \in (0, \infty)\} .$$

Let $\mathbb{R}^+(x)$ be the constant map from $\mathscr{X}$ to the single entity $\mathbb{R}^+ = (0, \infty)$, the set of positive real numbers; then

$$S(x) = \mathbb{R}^+(x)f(x : \cdot) .$$

The set $S(x)$ is called the likelihood function and can be labelled accordingly:

$$L(x: \cdot) = \mathbb{R}^+(x)f(x: \cdot) .$$

The problem of inference thus becomes *the analysis of* $L(x_0: \cdot)$ *with* C_c.

The likelihood function $L(x: \cdot)$ is all that a traditional model C_c presents concerning a response value x; different points x having the same likelihood function are not distinguished by the model. The use of the model thus effectively makes the reduction, $x \to L(x: \cdot)$, and replaces x_0 *with* C_c by $L(x_0: \cdot)$ *with* C_c. The traditional model C_c *does not* distinguish response values having the same likelihood. There is then no need or use for a theory of sufficiency together with a sufficiency principle to attain the weaker result—response values having the same likelihood *should not* be distinguished.

Consider the following argument L for a further reduction. Let A be an open set in $\mathscr{X}$ and suppose the diameter of A is small enough that $f(x: \theta)m(A)$ with x in A gives an adequate approximation to the probability content of A. And suppose further that $\mathscr{X}$ is partitioned into such sets A and a residual set F with $m(F) = 0$. Consider the observed x_0 with the model C_c. Let A be the set containing x_0. Then the observed x_0 with the model C_c asserts: an event A with probability $f(x_0: \theta)m(A)$ has occurred; the event $\bar{A}$ with probability $1 - f(x_0: \theta)m(A)$ has not occurred; and, accordingly, there is no observation on the conditional model $(1 - f(x_0: \theta)m(A))^{-1}f(x: \theta)$ on $\bar{A}$. With no observation on a traditional model there is no introduced information concerning θ_0. Thus x_0 with C_c asserts: an event A with probability $f(x_0: \theta)m(A)$ has occurred; and the event $\bar{A}$ with complementary probability has not occurred. Or equivalently: an event A with probability $f(x_0: \theta)m(A)$ has occurred. The arbitrariness of the diameter of A then produces the likelihood function. The problem of inference is then the analysis of the likelihood function $L(x_0: \cdot)$. Note that the particular likelihood function $L(x_0: \cdot)$ has probability of occurrence $L(x_0: \theta)$. Thus the *model* for the terminal entity $L(x_0: \cdot)$ is given by the entity itself. The primary step in the argument is concerned with no observation on the conditional model $\bar{A}$ with the result that the possible likelihood functions from $\bar{A}$ are omitted from further consideration. This takes the information about θ_0 from $L(x_0: \cdot)$ *with* C_c to $L(x_0: \cdot)$ *with* $L(x_0: \theta)$ and hence to $L(x_0: \cdot)$.

2. On the nature of information. Consider a physical system with a basic variable u having a known probability distribution. And suppose the system is enclosed by a mechanism (OPR) that observes, processes, and reports on the otherwise inaccessible variable u. Now consider a performance of the system yielding a realized value u_0. And suppose the OPR mechanism reports an item of information concerning the realized value u_0.

As a first case suppose an investigator *has* information concerning the behavior of the mechanism, specifically that it reports whether or not an event A occurs. And suppose that on the performance he *receives* the information that A has

occurred. The investigator has the behavior $I_A(\cdot)$ and he receives the report $I_A(u_0) = 1$ where I designates the indicator function. From the report he knows that u_0 is in the set A. And from the behavior he knows that each value in A would have produced the same report. The investigator can then describe the inaccessible u_0 by classical probabilities: the conditional probabilities for u given A. Of course if $P(A) = 0$ then a limiting operation is needed to justify the conditional probabilities.

As a second case, suppose the investigator has only the information that A has occurred. He then knows of course that the realized u_0 is in the set A. But he does *not* know that each value in A would have produced the report. In fact the mechanism might have the following behavior characteristics: for some $B \subset A$ the mechanism might observe $I_B(\cdot)$ and report $I_A(e_0) = 1$ if $I_B(e_0) = 1$ and $I_{\bar{B}}(e_0) = 1$ if $I_B(e_0) = 0$. The investigator obtains information, true information, but he does not know the observational information on which it is based and he cannot describe the inaccessible e_0 by classical probabilities.

This second case demonstrates that reported information without knowledge of its observational source does not support classical probabilities. In such cases reported information is not in the form of an *event* (as used in Fraser (1968) page 11) and by consequent indeterminacies shows that classical probabilities are not available. The distinction that information may be observational and support classical probabilities or may be non observational and not support classical probabilities seems to be unrecognized in applications. It clarifies certain problems and indeterminacies in the literature; for example, Freund (1965) and the familiar tactic of choosing the test of significance to yield a desired decision.

More generally, suppose an investigator receives an item of information concerning a realized value u_0 on a probability space and suppose he knows the manner in which the item of information is generated. He can then check whether the item of information is observational. For this let A be the set of possible realized values for which the item of information is true; he can assert that *one of the values in A has occurred.* And from the manner in which information is generated he must also be able to assert that *each of the values in A could have occurred* and *each of the values in A would have produced the same item of information.* If reported information is observational in this sense then classical conditional probabilities given A describe the inaccessible realized values.

A third case, a more substantial illustration, is given in Section 3.

3. Outcome identification by the variation-response model. Consider a system $\mathscr{E}$ under investigation: let u be the variable describing the various components of variation in the system, $\mathscr{U}$ be the variation space, $\mathscr{A} = \{A\}$ be the σ-algebra of permissible events A on $\mathscr{U}$ derived from the open sets of a metrizable topology, and m be an atom-free measure on $\mathscr{A}$; let x be the response variable, $\mathscr{X} = \mathscr{U}$ be the response space, and $\mathscr{A}$ be the σ-algebra of events on $\mathscr{X}$. A

variation-response model is a class $C_1 = \{f(u:\lambda)m(du): \lambda \in \Lambda\}$ of probability measures (presented by differentials) over $\mathscr{U}$ and a class $C_2 = \{\phi : \phi \in \Phi\}$ of homeomorphisms $x = \phi u$ of $\mathscr{U}$ onto $\mathscr{X}$.

As in Section 1 consider the assembled information concerning $\mathscr{E}$. A variation-response model (C_1, C_2) is *relevant* for $\mathscr{E}$ if the assembled information identifies sources of variation u with relevant model C_1 and identifies C_2 as the set of possible response productions from the variation u. A variation-response model (C_1, C_2) is *adequate* for $\mathscr{E}$ if it is relevant for $\mathscr{E}$ and if (C_1, C_2) is effectively all the assembled information concerning $\mathscr{E}$. Now let θ_0 be the combination (λ_0, ϕ_0) where λ_0 designates the distribution that describes the variation and ϕ_0 designates the transformation that presents the response. And let u_0 designate the realized variation in a performance of the system and x_0 designate the resulting observed response value. The problem of inference then is to determine from an adequate model (C_1, C_2) the information about $\theta_0 = (\lambda_0, \phi_0)$ provided by x_0 with (C_1, C_2).

Now with an adequate model (C_1, C_2) consider the information concerning the inaccessible variation u that is provided by a response value x. The model C_2 describes the inaccessible variation u as an element of

$$S_x = \{\phi^{-1}x : \phi \in \Phi\},$$

a set of values labelled by the transformations in Φ. This information concerning the realized variation is observational (Section 2) if each of the possible values for u would have produced the same information. A typical value in the set S_x is ${\phi_*}^{-1}x$ with ϕ_* in Φ. Such a value would produce the realized response $\phi_0{\phi_*}^{-1}x$ and the model C_2 would then describe the inaccessible variation as an element of

$$\{\phi^{-1}\phi_0{\phi_*}^{-1}x : \phi \in \Phi\}$$

which can be reexpressed as

$$\{h^{-1}x : h \in \phi_*{\phi_0}^{-1}\Phi\}.$$

The information is the same if the labelling class is the same; that is, if $\phi_*{\phi_0}^{-1}\Phi = \Phi$ for all ϕ_* in Φ, or if ${\phi_0}^{-1}\phi_*\Psi = \Psi$ for all ϕ_* where $\Psi = \{{\phi_0}^{-1}\phi : \phi \in \Phi\}$ is a class of transformations $\mathscr{U} \to \mathscr{U}$ that includes the identity, or if $\phi\Psi = \Psi$ for all ϕ in Ψ, or if Ψ is a *group* of transformations. Note that if Ψ is a group then it is independent of the ϕ_0 in its definition.

Thus a model C_2 produces observational information if and only if $\phi^{-1}\Phi$ is a transformation group on $\mathscr{U}$ or equivalently if and only if $\Phi\phi^{-1}$ is a transformation group on $\mathscr{X}$. In such a case the class $C_2 = \Phi$ can be factored as $\Phi = \phi_*\Psi$ where ϕ_* is an element of Φ and the variation-response model (C_1, C_2) is called a *group-map* model. Except for the response relabelling provided by the map this is the structural model examined in Fraser (1968).

For a group-map model (C_1, C_2) the response expressions C_2 provide observational information concerning the inaccessible variation; the information

identifies a set of possible values and gives no differential information within that set. In this sense the response expressions (group-map expressions) can be said to be *selectively transparent*.

For a group-map model (C_1, C_2) the initial probability model C_1 is changed by the response observation x_0 to a likelihood function for the identified set on the variation space *and* a conditional probability model describing the unidentified variation value in that set. The likelihood function with its model gives the information concerning λ_0 (as discussed in Section 1) and the conditional probability model, for any given λ, describes the only remaining unknown in the performance of the system.

4. Distribution identification with a group-map model. Consider a relevant group-map model (C_1, C_2) for a system $\mathscr{E}$. For ease of notation take an element ϕ_* in Φ and relabel the response space changing x to ${\phi_*}^{-1}x$. The class of transformations then becomes $\Psi = {\phi_*}^{-1}\Phi$ which is a group.

Now suppose that the distribution or class of distributions C_1 describing variation is not known and that a large number of observations on the system $\mathscr{E}$ is to be taken with the hope of identifying the distribution. Let $(u_1, \cdots, u_n)$ be the realizations for variation; then $(\phi_0 u_1, \cdots, \phi_0 u_n)$ would designate the response observations $(x_1, \cdots, x_n)$. The investigator can take any ϕ_* in Ψ and examine $(\phi_* x_1, \cdots, \phi_* x_n) = (\phi_* \phi_0 u_1 \cdots, \phi_* \phi_0 u_n)$ as a sample from a possible distribution for variation. Such samples and the corresponding distributions differ one from another by an element in the *group* Ψ. Thus with a sufficiently large number of performances the distribution describing variation can effectively be identified except for a transformation in the group Ψ.

Now consider this in a possibly different context, a context in which the sources of variation would not be altered but the level or form of the response could be different but within the range expressed by Ψ. Let u be one of the identified variables describing variation with its distribution given by $f(u)m(du)$, and let ϕ_0 designate the transformation that expresses the response. If the alternative variable $u^* = hu$ had been used to describe variation then $\phi_0{}^* = \phi_0 h^{-1}$ would be the transformation that expresses the same response. Of course ϕ_0 and $\phi_0{}^*$ present the same response. And in a partial sense they are the same transformation, they differ only due to the arbitrariness in the labelling of the sources of variation.

As an example, consider a response $x = \theta + u$ that is a general-level presentation of internal variation of known form—known from preceding experience with a similar system or with the same system in a modified context. The variable u could be located so that its central value is zero; the relocation θ then expresses explicitly the location of the response as measured by its central value. If the variable $u^* = k + u$ had been used then the response would be expressed by the relocation $\theta - k$. In this alternative case the transformation $\theta - k$ would not immediately express numerically the general response level.

Thus: the identification of the distribution describing variation leaves an arbitrary element, a transformation in the group Ψ. Consequently the transformation that expresses the response involves an arbitrary element, right multiplication by a transformation in the group Ψ. The significance of a transformation expressing the response lies in the characteristics of the response produced; two transformations differing by right multiplication correspond to two different labellings of the sources of variation—truly an arbitrary thing.

5. The likelihood approach. The likelihood approach to a system $\mathscr{E}$ commences with a relevant model—a relevant traditional model C; for convenience consider here the continuous case $C_{c\cdot\varepsilon}$. The likelihood approach to inference is the analysis and presentation of the set $L(x_0: \cdot)$.

The proponent of the likelihood approach accepts the reduction from *analysis of* x_0 *with* C_c to *analysis of* $L(x_0: \cdot)$ *with* C_c; he accepts the reduction not on the analysis in Section 1 but typically on other grounds such as a sufficiency principle. And the proponent accepts the further reduction from *analysis of* $L(x_0: \cdot)$ *with* C_c to *analysis of* $L(x_0: \cdot)$ *alone*; he accepts this reduction on principle, the likelihood principle or by argument from a conditionality principle. He might welcome the argument L that $L(x_0: \cdot)$ is all that the model provides concerning the unknown θ_0. The proponent of the likelihood approach typically *believes* that all information concerning θ_0 is contained in the likelihood $L(x_0: \cdot)$.

The likelihood approach produces the summary inference

$$L(x_0, C_c) = L(x_0: \cdot) = \mathbb{R}^+(x_0) f(x_0: \cdot)$$

concerning the unknown θ_0. For simplicity of expression it may extract a representative from $L(x_0: \cdot)$ with maximum value 1 or with integrated value 1 relative to some convenient measure for θ:

$$L_1(x_0, C_c) = f(x_0: \cdot)/\sup_\theta f(x_0: \theta) ;$$
$$L_2(x_0, C_c) = f(x_0: \cdot)/\int f(x_0: \theta)\, dn(\theta) .$$

A more liberal form of likelihood would allow a nonnegative weight function $w(\cdot)$ expressing in some manner an assessment from outside $\mathscr{E}$ concerning θ_0, and would produce a composite assessment

$$L_3(x_0, C_c, W) = \mathbb{R}^+(x_0) w(\cdot) f(x_0: \cdot) .$$

The Bayesian feels that the likelihood approach is acceptable as far as it goes but it does not go far enough. By his persuasion the Bayesian does not know how to make inferences from just a likelihood function; rather he treats the likelihood function as a device that modifies his numerically expressed preferences from before the performance to give his preferences concerning θ_0 after the performance; he treats the likelihood function as a transition function that operates at the time of the performance. The Bayesian believes that all unknowns can and should be assessed by a numerical expression of subjective

feelings and preferences, so do not stop with the likelihood function, use it as a transition function.

The proponent of the structural approach believes that an adequate statistical model should always be used. If the relevant traditional model for $\mathscr{E}$ is adequate, then the structural proponent would accept the reduction to *analysis of* $L(x_0: \cdot)$ *with* C_c, and would give consideration to the argument L for further reduction to *analysis of* $L(x_0: \cdot)$. He accepts the general principle germain to the likelihood approach that inference should be based on the realized response x_0 and the assembled information concerning $\mathscr{E}$ (the possibilities for what could be as opposed to what could not be) and should exclude feelings and preferences that attempt to provide shadings one to another among the possibilities; to this extent he is in accord with the likelihood approach. The structural proponent, however, views the traditional model as being a severe abstraction, an abstraction that generally neglects many characteristics of $\mathscr{E}$ that may be relevant to identifying θ_0: such seemingly minor things as distance and differentiability on $\mathscr{X}$ and on Ω; and such seemingly central things as where and how does variation in the response originate, in other words, what is the basic probability space. Even the minor things invalidate argument L. The structural proponent believes that in most cases the traditional model is inadequate for the system $\mathscr{E}$ being modeled and accordingly believes, in those cases, that it is inappropriate as a basis for inference: *substantiated information has been omitted from the assembled information concerning the system* $\mathscr{E}$ *under investigation.*

6. The bayesian approach. The bayesian approach to a system $\mathscr{E}$ commences with a relevant model, a relevant traditional model C; for convenience, consider here the continuous case C_c. And the approach has a *prior* model

$$P = \{p(\theta)n(d\theta) : p \in \Pi\},$$

one or several distributions suggesting a random source $\mathscr{H}$ for the unknown θ_0 in the system.

The bayesian approach is not directly involved with cases where there *is* a random system $\mathscr{H}$ with model P. In such cases the system under examination is $(\mathscr{H}, \mathscr{E})$ and the model is (P, C_c); as examples, tests on the progeny from a mating of parents with known genetic characteristics; "empirical Bayes" where sufficient data are available to support a model P.

The bayesian approach introduces the hypothetical system $\mathscr{H}$ to suggest the differential feelings and beliefs that the investigator has concerning the possible values for θ_0; or to suggest some kind of indifference among values for θ_0; or to suggest properties that, conveniently, are functionally compatible with the model C.

The bayesian approach accepts the composite model (P, C_c) as appropriate to the analysis of x_0 with C_c. The composite model has probability differential $p(\theta)f(x: \theta)n(d\theta)m(dx)$. The observed x_0 gives observational information on

(θ_0, x_0) and changes the preceding probability distribution to the conditional distribution given x_0,

$$k_p^{-1}(x_0)p(\theta)f(x_0 : \theta)n(d\theta) ,$$

as the description of the inaccessible θ_0. Note that this conditional distribution can be calculated from $p(\cdot)$ and $L(x_0 : \cdot)$ alone, by modulating one by the other.

The bayesian approach produces the summary inference

$$B(x_0, C_c, P) = \{k_p^{-1}(x_0)p(\theta)f(x_0 : \theta)n(d\theta) : p \in \Pi\}$$

concerning the unknown θ_0, where the indexing class Π is to suggest the feelings, the indifference, or the compatibility.[1]

A radical form of the bayesian approach has allowed the prior model P to depend on the observed response, that is, P becomes P_{x_0}. In general the composition (P_x, C_c) is not a probability model and there is, accordingly, no conditional distribution. However, by formal analogy with the simpler case the following radical inference is obtained

$$B^*(x_0 : C_c, P_{x_0}) = \{k_p^{-1}(x_0)p(\theta)f(x_0 : \theta)n(d\theta) : p \in P_{x_0}\} ;$$

an example may be found in Box and Cox (1964).

The proponent of the likelihood approach is against the use of subjective, personalistic, or indifference probabilities as embodied in a model P or P_x—he is against the introduction of a hypothetical system $\mathscr{H}$. His purpose is to investigate the system $\mathscr{E}$ and to extract information concerning the value of θ_0 and he believes this scientific activity should be kept separate from diffuse feelings and preferences concerning the value of θ_0. The proponent of liberal likelihood acknowledges that he can obtain any bayesian result by the use of measures and weight functions but he does not call them probabilities and in turn has some reassurance that they will not be treated as ordinary probabilities. The proponent of likelihood believes that his inferences are specific to $\mathscr{E}$; but of course if some client wants a basis for action and feels a need to incorporate diffuse feelings and preferences then certainly an option for that client is to represent these by measures or weight functions and incorporate them with likelihood, whatever the result may mean. He does believe that such client expedients should be quite separate from the scientific activity for $\mathscr{E}$ of extracting and presenting information concerning θ_0.

The proponent of the structural approach believes that an adequate statistical model should always be used. If there is an adequate statistical model that incorporates information additional to some relevant traditional model, then the bayesian use of just likelihood in fact suppresses this additional information; which poses greater needs for justification of the bayesian premises. An adequate statistical model effectively presents the possibilities for what could be (as opposed to what could not be) all within a reasonable degree of approximation; the bayesian approach purports to add feelings and preferences to provide

shadings one to another among the possibilities. The structural proponent believes that the Bayesian abuses scientific principle by prescribing that such feelings, indifference, and compatibility be input to analysis and thereby superimposed on results from the system $\mathscr{E}$. He feels that progress in science has been based on clearly separating feelings and impressions from observation, experimentation, and the inference process; for history records how biased these feelings and impressions can be and large portions of experimental design exist for the prime purpose of excluding effects from such feelings and impressions. This is not to suggest that ideas for new investigations should not be influenced by the scientist's feelings and impressions, although many may be wrong for one close, but rather that the investigation should be an entirely separate activity. The structural proponent believes that the Bayesian never knows the results from his experiment or system $\mathscr{E}$—only a compound of these with feelings and impressions: *hypothetical information has been added to the assembled information concerning the system $\mathscr{E}$ under investigation.*

But even more fundamental than whether one would want probabilities describing unknowns based on information from diffuse sources is the question of whether one can have probabilities in such situations. The Bayesian argues yes from plausible premises often in a gambling context. The results from Section 2, however, show otherwise—probabilities are not generally available to describe unknowns on the basis of information.

7. The structural approach. The structural approach to a system $\mathscr{E}$ commences with an adequate statistical model. The underlying principle is that inference for a system $\mathscr{E}$ should be based on observation and the assembled information as to what could be as opposed to what could not be—all within a reasonable degree of approximation. If the adequate model is traditional then the approach leads (Section 1) to the likelihood function together with the statistical model. If the adequate model is traditional but with additional properties such as distance and continuity then the reduction to likelihood may not be justified and the approach opens on the range of classical methods such as confidence intervals and significance tests. If the adequate model is variation-response of the group-map form, then part of the inference process is predicated by probability theory itself while the other part falls to likelihood and the other classical methods (Section 3 and Fraser (1968)). If the adequate model is a general variation-response model then the approach opens many theoretical questions touching on the classical methods and the determination of likelihood itself; see Fraser (1972).

The likelihood proponent focuses his attention on the group-map model with no distribution parameter λ. As part of the structural solution he notes a derived distribution describing the unknown θ and sees that it is just a likelihood function. He would concede that it has likelihood form only with respect to a particular measure. He does feel, however, that the probability argument to

obtain the derived distribution is a long route to a likelihood function. As a believer in likelihood he feels there is no need for a long route to a likelihood function when one can, in fact, start with the likelihood function.

The Bayesian also focusses on the derived distribution in the group map case, and feels that the probability arguments to obtain the derived distribution form a long route to a distribution that is often in $B(x_0, C_c, P)$ and is sometimes the single distribution in $B(x_0, C_c, P)$—provided a customary convention is used to construct P (cf. Section 8). He feels that the derived distribution is just one of many provided by the bayesian approach, so why such a long argument to obtain so little.

The proponent of the structural approach would reply that if only answers are important then the easiest approach is to commence with all answers in any way suggested by intuitive considerations. By contrast the structural approach is primarily concerned with finding *the* answer, with finding what is predicated by observation and assembled information.

From a pragmatic point of view one can examine cases where all three methods are available. To get substance one needs to go to variation-response models and to avoid open issues restrict attention to group-map models. Consider $f(\phi^{-1}x : \lambda) = g(\phi^{-1}[x] : D, \lambda)h(D : \lambda)$ where $[x]$ is a transformation variable for the group Φ and $D = D(x) = [x]^{-1}x$ designates the orbit (see Fraser (1968), (1972)); the underlying group-map model is $x = \phi u$ where u has distribution $f(u : \lambda)$.

The structural approach gives, *among other results*, the distribution $k(x)g(\phi^{-1}[x] : D, \lambda)d\nu(\theta)$ for ϕ given λ and the likelihood $\mathbb{R}^+(D)h(D : \lambda)$ for λ.

The likelihood approach gives $\mathbb{R}^+(x)g(\phi^{-1}[x] : D, \lambda)h(D : \lambda)$. Extended likelihood methods can then give the sectional likelihood $\mathbb{R}^+(x)g(\phi^{-1}[x] : D, \lambda)$ for ϕ given λ and the profile likelihood $\mathbb{R}^+(x)g(\hat{m}_\lambda(D) : D, \lambda)h(D : \lambda)$ for λ ($\hat{m}$ maximizes g under first argument variation). The sectional likelihood if given an appropriate measure reproduces the distribution for ϕ given λ. On the other hand the profile likelihood is typically not even a likelihood, is not the likelihood from the input variable D but has an extraneous and improper factor $g(\hat{m}(D) : D, \lambda)$.

The bayesian approach typically needs the radical form of analysis. A conventional distribution for ϕ given λ can produce the distribution for ϕ given λ obtained from the structural approach. But to obtain the component $h(D : \lambda)$ for λ, the prior for ϕ typically would need to depend on the observation x and be chosen in a goal-directed manner to eliminate the factor $g(\phi^{-1}[x] : D, \lambda)$. Such a procedure might have some reassurance if the goal were known but the goal (obtained deductively) seems exclusively a result of the structural approach.

8. Structural support for the Bayesian. Consider a group-map model (C_1, C_2) and suppose that Λ has a single element, that $\mathscr{X}$ has been relabelled (Section 4) so that the class of transformations Ψ is a group, and that m is an invariant measure on $\mathscr{X}$:

$$C_1 = \{f(u)m(du)\}\,, \qquad C_2 = \{\phi : \phi \in \Psi\}\,.$$

The corresponding traditional model describing the response variable x is

$$C = \{f(\phi^{-1}x)m(dx) : \phi \in \Psi\}\,.$$

This model has invariant form under the group Ψ: the variable x with distribution labelled ϕ in Ψ is transformed by an h in Ψ into the variable hx with distribution $h\phi$ in Ψ. The transformation h *changes* the variable (from x to $y = hx$) and it *changes* the parameter value labelling the response distribution (from ϕ_0 to $\psi_0 = h\phi_0$); but it *does not change* the model.

Now consider a bayesian analysis based on arguments of indifference. A transformation h carries x into $y = hx$ and ϕ into $\psi = h\phi$; a prior differential $p(\phi)n(d\phi)$ is carried in to $p(h^{-1}\psi)n(dh^{-1}\psi)$. The model is unchanged; a Bayesian could argue that the prior distribution should also remain unchanged:

$$p(h^{-1}\psi)n(dh^{-1}\psi) = p(\psi)n(d\psi)\,.$$

This implies that $p(\phi)n(d\phi) = c\mu(d\phi)$ where μ is the left Haar measure on Ψ; typically this is an improper distribution but it can be normalized on a large but compact set. The resulting posterior distribution for ϕ is

$$k(x_0)f(\phi^{-1}x)\mu(d\phi)\,.$$

Surprisingly, however, the Bayesians seem to prefer a prior measure that is right Haar (Jeffreys (1961); Stone (1965)). For the group-map model (C_1, C_2) arguments to support this choice of right Haar measure are available from the structural approach—given of course the premise of an indifference prior.

Consider the results from Section 4 concerning the identification of the variation distribution. Let u now be one of the identified variables describing variation and ϕ_0 be the corresponding transformation giving the response. And let h be an element in the group Ψ. Then $u^* = hu$ is an alternative variable describing variation and $\psi_0^* = \psi_0 h^{-1}$ is the corresponding transformation giving the response. An indifference argument relative to the group Ψ should reflect the arbitrariness in the choice of representative for response expression from variation:

$$p(\phi h)n(d\phi h) = p(\phi)n(d\phi)\,.$$

This implies that $p(\phi)n(d\phi) = c\nu(d\phi)$ where ν is the right Haar measure. The resulting posterior is

$$k(x_0)f(\phi^{-1}x)\nu(d\phi)\,.$$

In this analysis the variable, the model and the essential parameter that describes the response distribution remain unchanged; the indifference argument expresses the *arbitrariness* in the group element used within the model to stand for the essential parameter.

The structural examination of the group-map model leads to an appropriate

indifference argument: no change of variable, parameter, model; indifference to choice of representative for response expression. This is not presented in support of the bayesian approach, but as an indication of how structural argument can identify an appropriate entity otherwise undetermined in the bayesian method.

9. Probabilities for unknowns. Probabilities obtained from assembled information about a system have two distinct roles. One is the description or prediction of what will happen in a future performance of the system. The other is the description of a past realization of the system, a realization that typically would be partially or completely concealed to avoid trivial descriptions. The kind of information that supports these latter probabilities was examined in Section 2; the effect of proper information is to change initial probabilities into conditional probabilities—given the observational content of the information.

Consider the second role for probabilities: the description of a partially or completely concealed realization of a stable system or more generally, perhaps, the description of an unknown in the physical environment. With observational information on a stable system probabilities are available (Section 2). More generally with frequency, observational and other information including feelings and impressions, probabilities are *claimed* by the Bayesians to be still available; the results of Section 2, however, show that this general bayesian claim is false. The question then remains: do probabilities for unknowns extend at all beyond the strict case satisfying Section 2. In this section attention is restricted to contexts satisfying Section 2 but unknowns other than the realization itself are considered.

There are of course some statisticians of extremely conservative persuasion who are reluctant to use probability even to describe a realized value from a stable system. In being so, they are of course denying the raison d'être for conditional probability and hence denying a substantial part of probability theory. Most statisticians however accept this use of probability and the arguments for it are not presented here.

There may then be some of conservative persuasion who are reluctant to use probability as based on observational information from the OPR mechanism of Section 2; they must however face the substantial arguments that there is still no differential information among the possibilities identified by the information. Again, the arguments for this use of probability are not presented here.

For exemplification consider a measuring instrument with known stable behavior and suppose two measurements are made on an unknown physical constant. To keep the details simple suppose the two uses of the instrument are represented by one toss of two symmetrical dice and suppose the values obtained are added to the physical constant to give the two reported measurements.

Let u_1 and u_2 designate the results for the first and the second die, let θ_0 be the physical constant, and let $x_1 = \theta_0 + u_1$ and $x_2 = \theta_0 + u_2$ be the reported

measurements; suppose $x_1 = 126$ and $x_2 = 129$.

The OPR conditions for observational information are satisfied. The observational information is

$$u_2 - u_1 = x_2 - x_1 = 3 ;$$

and there is no information on the location of (u_1, u_2) as given by, say, u_1. The observational information changes the probabilities 1/36 for each possible pair into probabilities $\frac{1}{3}$ for each of the pairs (1, 4), (2, 5), (3, 6). These are classical probabilities describing the three possible physical states for the entity before the investigator. Generally, statisticians who acknowledge conditional probability would acknowledge these probabilities. The acute investigator then notes that the label $\theta_0 = 125$ attaches to the physical state (1, 4), the label $\theta_0 = 124$ to the physical state (2, 5), and the label $\theta_0 = 123$ to the physical state (3, 6). The probabilities describing the possible physical states are then in fact saying $\Pr(\theta_0 = 125) = \frac{1}{3}$, $\Pr(\theta_0 = 124) = \frac{1}{3}$, and $\Pr(\theta_0 = 123) = \frac{1}{3}$. The student, the scientist, the experienced gambler—certainly the experienced gambler—would be prepared to bet on the concealed dice and he would use the probabilities recorded above. In doing so he is actually betting on the θ_0 value, as he would readily admit: for the true value of θ_0 is in fact the one that labels the realized pair of values on the dice. Probabilities are describing things in the physical world and we must accept the consequences of those descriptions.

Classical probabilities for physical states can produce classical probabilities for entities within the observational system that provides information. This happens generally with the group-map models of Section 3.

Of course a committed Bayesian might say that with a pattern of possible θ_0 values the probabilities would be different. But such a source pattern for the θ_0 values was not part of the premises in Section 2, not part of the present analysis, and is not part of typical scientific analyses.

REFERENCES

[1] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B.* **26** 211-243.

[2] Fraser, D. A. S. (1968). *The structure of Inference.* Wiley, New York.

[3] Fraser, D. A. S. (1972). The determination of likelihood and the transformed regression model. *Ann. Math. Statist.* **43** 898-916.

[4] Freund, J. E. (1965). Puzzle or paradex. *The American Statistician* **20** 4, 44.

[5] Jeffreys, Sir Harold (1961). *Theory of Probability.* Oxford Univ. Press.

[6] Stone, M. (1965). Right Harr measure for convergence in probability to quasi posterior distributions. *Ann. Math. Statist.* **36** 440-53.