The Canadian Journal of Statistics
Vol. 14, No. 3, 1986, Pages 181–199
La Revue Canadienne de Statistique

On principles and arguments to likelihood

Michael J. EVANS[1], Donald A.S. FRASER[2], and Georges MONETTE[3]

University of Toronto[1,2], *York University*[2,3] and *University of Waterloo*[2]

Key words and phrases: Principles, conditionality, sufficiency, likelihood, content, evidential meaning.
AMS 1980 subject classification: Primary 62A99.

ABSTRACT

Birnbaum (1962a) argued that the conditionality principle (**C**) and the sufficiency principle (**S**) implied the likelihood principle (**L**); he then argued (Birnbaum 1972) that **C** and a mathematical equivalence principle **M** implied **L**. Evans, Fraser, and Monette (1985a) gave reference details, and this paper gives proof that **C** alone implies **L**. The level of support by the profession for **L** is sharply less than that for **S** or even for **C**; thus the paradoxical nature of these results. In this regard, we elaborate on the Monette example (Fraser, Monette, and Ng 1984), which provides a strong case against **L**. We also examine closely the various proofs linking the principles and find that **S** and **C** can each be used operationally to suppress information otherwise deemed relevant. From another viewpoint this says that **S** and **C** can each be used in contexts that directly conflict with the original examples and motivations supporting them; the principles can thus be viewed as inappropriately used, or more strongly, as invalid. In either case, the result that **C** and **S** imply **L** or that **C** implies **L** can be regarded as noneffective in the context of discriminating applications. A resolution of the apparent anomalies can be obtained by allowing the statistical model to include ingredients additional to those usually present (particularly for subsequent use with conditionality), or alternatively by restricting the application of the principles to contexts where the conflicts would seem not to arise.

RÉSUMÉ

Birnbaum (1962a) a montré que le principe de vraisemblance (**L**) découle du principe de conditionnement (**C**) et du principe d'exhaustivité (**S**). Plus récemment, Birnbaum (1972) a démontré que **L** découle de **C** et d'un certain principe d'équivalence mathématique (**M**). On peut trouver un exposé de ces résultats dans l'article de Evans, Fraser et Monette (1985a). Dans le présent travail, on démontre que **L** est une conséquence de **C** seulement. Tous ces résultats sont paradoxaux à cause du peu de faveur dont jouit **L** dans la profession, à comparer à **S** ou même à **C**. A ce propos, nous reprenons un exemple de Monette (Fraser, Monette et Ng 1984) qui plaide en défaveur de **L**. Un examen attentif des démonstrations qui unissent ces différents principes nous amène à conclure que **S** et **C** peuvent tous deux être utilisés en pratique pour supprimer de l'information précieuse. Vu d'un autre angle, ceci revient à dire que ces deux principes peuvent être invoqués dans des contextes qui entrent en conflit direct avec les exemples originaux et leur motivation. On peut donc en venir à rejeter **S** et **C**, ou tout au moins à juger qu'ils sont employés à mauvais escient. Dans les deux cas, dire que **S** et **C** entraînent **L** ou que **C** entraîne **L** peut être considéré comme non avenu dans le cadre d'applications choisies. On peut réconcilier ces anomalies apparentes en incluant certains ingrédients nouveaux dans le modèle statistique usuel, surtout si on veut faire intervenir le principe de conditionnement. Une autre possibilité consiste à restreindre l'application de ces principes aux contextes où les conflits ne semblent pas surgir.

1. INTRODUCTION

We are concerned here with the general role of principles in statistical inference and more specifically with implications among conditionality, sufficiency, mathhematical equivalence, and likelihood as considered in Birnbaum (1962a, 1972) and in Evans, Fraser, Monette (1985a).

Within the context of what can be called classical frequency-based statistical inference, Birnbaum (1962a) argued that the conditionality and sufficiency principles imply the likelihood principle. Birnbaum (1972) subsequently argued that the conditionality principle and a mathematical equivalence principle imply the likelihood principle. The wide acceptance of the conditionality, sufficiency, and mathematical equivalence principles and a general rejection of the likelihood principle caused these results to be regarded as disturbing paradoxes within the foundations of statistical inference and within the interpretations of the statistical process. The results point to a clear need for an incisive assessment of the use of statistical principles in order that the paradoxes can be understood and the difficulties avoided.

Doubts have been expressed concerning Birnbaum's results; for example, Fraser (1963), Durbin (1970), Kalbfleisch (1975), and Joshi (1976). None of these counterarguments can, however, be regarded as providing definitive grounds for rejecting the results. Other discussion of Birnbaum's results may be found in Birnbaum (1962b), Hartigan (1967), Hajek (1967), Birnbaum (1970), Basu (1975), Dawid (1977), Godambe (1979), Barnard and Godambe (1982), and Berger and Wolpert (1984).

In this paper we present the further result that the conditionality principle alone implies the likelihood principle. This result is developed from the material on cross-embedded models in Evans, Fraser, and Monette (1985a).

The difference in the professional support for the conditionality as opposed to the likelihood principle poses the central contradiction or paradox. Does direct support of conditionality imply support of the likelihood principle? The answer is yes—but only in he context of the standard statistical model, the ordinary formulation of the principles, and their free and uncritical use. The details of the proofs in fact show that a distribution which ostensibly contains no information can be used to suppress information beyond that in the likelihood funtion, and that this happens precisely because the ordinary statistical model contains only the space-algebra-density ingredients. Alternatively, the proofs can be viewed as inappropriate, as they involve uses of sufficiency and conditionality in contexts where the applications of these principles are operationally in conflict with or in contradiction to their original supporting examples. Accordingly, the usual justifications for these principles are not then available in these contexts.

Possible resolutions of the problems posed by these results lie in two directions: acknowledging that the statistical model should contain additional elements, particularly for its use with forms of conditionality; or revising the formulation of the principles of sufficiency and conditionality to correct for the context invalidities indicated by the proofs.

2. THE INFERENCE BASE AND ITS CONTENT

In the pattern of classical frequency-based statistical inference, we follow Birnbaum (1962a) and others and work from the *assumption* that a statistical model M has the form

$$M = (S, \mathbf{A}, \{f_{\theta} : \theta \in \Omega\}; \mu) \tag{2.1}$$

where S is a set, A is a σ-algebra on S, and $\{f_\theta : \theta \in \Omega\}$ is a class of probability densities with respect to a support measure μ on (S, A); for a valid application one value of θ is the true value.

As discussed in Fraser (1979), a valid statistical model for an application is descriptive and exhaustive. This reflects in two fairly obvious but different ways on the assumption concerning the model M.

The first has to do with a model being descriptive. As noted by Joshi (1976) and others, allowing such a broad mathematical generality for M can produce technical difficulties for some of the results in Birnbaum (1962a, 1972). However, if we place mathematical restrictions on the components of M, these difficulties can be avoided. The interested reader can consult Evans, Fraser, and Monette (1985b). The restrictions are founded on structural aspects of virtually all applications, and use background material from Rudin (1974). The mathematical restrictions are such that arbitrariness is avoided in the definition of the probability density function and will be implicitly assumed here without further comment. We note, however, that when S and Ω are finite the difficulties do not arise, and that this class of models is rich enough to illustrate all points raised in the paper. Accordingly we generally restrict our present analysis, and take A to be the power set on S and μ to be counting measure.

The second has to do with a model being exhaustive. The *physically* restrictive nature of the model (2.1) involving just a space, an algebra, and a class of densities can prevent full modelling in an application. This has been noted in Fraser (1968), and some additions to the model for specific contexts have been discussed in Fraser (1979). We will see that the restrictive nature of the model (2.1) is central to resolving the paradoxes relating conditionality and sufficiency to likelihood.

The restriction to models of the form (2.1) has been viewed as a *distribution principle* of statistical inference by Dawid (1977) and Godambe (1979). By contrast, however, we view it as concerning the *given*, not the process (statistical inference) from *given* to *conclusion*.

Now consider the model M together with a data value $s \in S$. We form the model-data combination $I = (M, s)$, called an *inference base* in Fraser (1979) and referred to as "an instance of statistical evidence" in Birnbaum (1962a). Note that when we consider an inference base we are assuming that the model M is a valid model (Fraser 1979) for an application, and that the data value was the observational material from that application.

Birnbaum (1962a, 1972) restricted his attention to the class $\mathcal{I}(\Omega)$ of all inference bases with parameter space Ω, and we continue with this restriction. A more general consideration with various Ω is not needed for our present discussion of the arguments from conditionality and sufficiency to likelihood, but is discussed briefly at the beginning of Section 2.

In relation to applications we note that $\mathcal{I}(\Omega)$ operationally contains many identical copies of any particular inference base I, corresponding to various applications in various contexts. These various copies are of course identical and are represented formally by just a single element I in $\mathcal{I}(\Omega)$. In the applications, however, they would typically correspond to various true values for the parameter in Ω.

We note, parenthetically, that at a crucial step in Birnbaum's presentation of a mixture model two different inference bases must refer to the same true value for the parameter. This technically could be a difficulty for Birnbaum's argument. Of course there is no intention that each inference base in $\mathcal{I}(\Omega)$ should refer to the same true value. The solution to the technical difficulty lies in considering identical copies of any inference base and at the crucial step considering copies that in context refer to the same true θ; this point is discussed later.

Many approaches to statistics involve additional ingredients beyond the inference base, such as prior distributions, classes of decision functions, loss functions. We view such components as *additives* to the inference base and stress that our discussion here does not involve these elements. Discussion concerning the relevance of such additives to inference can be found in Brenner, Fraser, and Monette (1981).

Birnbaum (1962a) introduced the notation $\text{Ev}(M, s)$ for "the evidential meaning of a specified instance, (M, s), of statistical evidence". In Birnbaum's analysis the function Ev was used only to induce the equivalence relation defined by as follows: $(M, s) \sim (M', s')$ if and only if $\text{Ev}(M, s) = \text{Ev}(M', s')$.

The ultimate goal for a theory of inference in the context of $\mathcal{I}(\Omega)$ is to express the fundamental information *content* of an inference base I in $\mathcal{I}(\Omega)$, in other words, what the model and data in $I = (M, s)$ say concerning the unknown θ in Ω. Accordingly, we let $\text{cont}(M, s) = \text{cont}(I)$ designate the collection of all logical implications and constraints concerning the true θ as provided by the inference base $I = (M, s)$. From this viewpoint the ultimate goal of statistical inference is the clear, accessible presentation of $\text{cont}(I)$ for each I in $\mathcal{I}(\Omega)$. While we do not attempt to determine cont in this way, we do target on a first goal of establishing under what conditions $\text{cont}(I_1) = \text{cont}(I_2)$, that is, on determining the preimage partition of the function cont on $\mathcal{I}(\Omega)$. The shift in emphasis from meaning to logical content underlies the change in notation from Ev to cont.

3. RELABELLING

We briefly mention some relabelling issues that arise for inference bases $I = (M, s)$ and for a class $\mathcal{I}(\Omega)$ of all inference bases with a given parameter space Ω.

First we note that two inference bases, $I_1 \in \mathcal{I}(\Omega_1)$ and $I_2 \in \mathcal{I}(\Omega_2)$, that are identical under a bijection $g : \Omega_1 \rightarrow \Omega_2$ can in context correspond to the same physical reality, the difference being only in the labels used to refer to the possible characteristics for the physical reality. The equivalence of such inference bases would thus assume that the labelling is nonsubstantive. However, as we have defined the model in Section 2, we have that I_1 is different from I_2 (unless g is the identity function). We do not further address such relabelling issues, but mention an invariance principle in Hartigan (1967) and the use of equivalence classes to reduce the model in Fraser (1979).

Now consider relabelling on the sample spaces. For two inference bases $I_1 = (M_1, s_1)$ and $I_2 = (M_2, s_2)$, suppose there is a measureable bijection $g : S_1 \rightarrow S_2$ such that $f_\theta^1(s) = f_\theta^2(gs)$ for all s and θ. We note that in a suitable context the two inference bases could be in correspondence with the same physical reality, the difference being only in the labels attached to possible outcome values. With our present model formulation the inference bases I_1 and I_2 are, however, different, except in the trivial identity case.

Consider the class $\mathcal{I}(\Omega)$, and let **R** designate a *relabelling principle* that asserts $\text{cont}(I_1) = \text{cont}(I_2)$ if the bijection in the preceding paragraph holds. The use of this principle establishes an equivalence relation on $I(\Omega)$ and removes the dependence on the sample-space labelling. We will refer to this principle later in our discussion of the basic statistical principles of sufficiency, conditionality, and likelihood. The principle **R** has been discussed in Godambe (1979) (with designation $\bar{M}$), and from a somewhat different viewpoint in Fraser (1979, pp. 79f).

4. THE STATISTICAL PRINCIPLES

While the equivalence relation Ev is not formally specified by Birnbaum, he does consider a number of possible properties of Ev that relate to principles that various groups of statisticians employ. These are the principles of conditionality, sufficiency, mathematical equivalence, and likelihood. We record these principles from Birnbaum (1972), with slight rephrasing, and use the notation "cont" discussed in Section 2 to emphasize our shift in emphasis from *meaning* to *content*.

4.1. Likelihood Principle.

For an inference base $I = (M, s)$ the *likelihood function* is given by

$$\text{lik}(I) = \{kf\cdot(s) : k \in \mathbb{R}^+\},$$

the class of positive multiples of the density function evaluated at the data point s and treated as a function of the parameter. Note that lik(I) is formally a ray from the origin in the vector space $\mathbb{R}^\Omega$.

The *likelihood principle* is then given as

$$\mathbf{L}: \quad \text{cont}(I_1) = \text{cont}(I_2) \quad \text{if } \text{lik}(I_1) = \text{lik}(I_2).$$

The likelihood principle treats all information beyond the likelihood function as irrelevant information for inference concerning θ. Some support for the likelihood principle exists among groups of statisticians, but some of those who support it also work in areas of statistics precluded by the likelihood principle.

4.2. Sufficiency Principle.

For a model M, the map $t: (S,A) \rightarrow (T,B)$ is *sufficient* if the conditional model induced by t has only one distribution for each value of t. For the more general circumstances involving nondiscrete models we assume the further restrictions on the function t as discussed in Evans, Fraser, and Monette (1985b).

The *sufficiency principle* is given as follows:

S: cont(I_1) = cont(I_2) if I_2 can be obtained from I_1 via a sufficiency map.

The sufficiency principle treats the information in the θ-free conditional model and its data as being irrelevant for inference concerning θ. As one resolution of the paradoxes concerning the principles we will examine a modification of this principle in Section 8.

4.3. Conditionality Principle.

For a model M, the map $u : (S,A) \rightarrow (U,B)$ is *ancillary* if the marginal model of u is θ-free. For the more general contexts we assume the further restrictions on u developed in Evans, Fraser, and Monette (1985b).

The *conditionality principle* is given as follows:

C: cont(I_1) = cont(I_2) if I_2 is the conditional inference base given the value of an ancillary for I_1.

The conditionality principle treats the information in the marginal model for the ancillary and in the collection of conditional models, other than the one indicated by the value of the ancillary, as being irrelevant for inference concerning θ.

A number of appealing examples exhibit this ancillarity property, perhaps the most prominent being that involving a random choice of measuring instrument (Cox 1958, Pratt 1961). For consider a random $(\frac{1}{2}, \frac{1}{2})$ choice among two measuring instruments (Inst_1, Inst_2) for measuring a parameter θ; the instruments support different statistical models. Persuasively it seems that the measuring instrument *actually used* should provide the statistical model for the inference concerning θ. For a discussion of such appealing examples for conditionality, see Fraser (1979, §3-2).

These special examples provide the grounds for the principle **C** in the general context of $I(\Omega)$. As part of a resolution of the paradoxes we will examine a modification of this principle in Section 8.

4.4. Mathematical Equivalence

Birnbaum (1972) inroduced a principle of *mathematical equivalence* which functions as a weakened version of sufficiency:

M: $\text{cont}(I_1) = \text{cont}(I_2)$ if $I_1 = (M, s_1)$, $I_2 = (M, s_2)$, and $f_\theta(s_1) = f_\theta(s_2)$ for all θ.

Given that the points s_1 and s_2 are identical within the model, except for labelling, then the information as to which of s_1 and s_2 has occurred is viewed as irrelevant for inference about θ.

Clearly the relabelling principle **R** implies **M**, for all we need is a mapping from S to S that is an identity except for the interchange of the points s_1 and s_2. A minor technical point arises however, as the relabelling principle **R** nominally refers to different inference bases I_1 and I_2 rather than a "change of sample point" within an inference base. This is directly handled by working with a copy of say $I_2 = (M_2, s_2)$ as discussed in Section 2. We also note that the conditionality principle **C** implies both **M** and **R**, as a consequence of Theorem 1 in Section 6.

4.5. A Note on implications among Principles.

Birnbaum viewed the principles **L**, **S**, **C**, as relations on $I(\Omega)$, that is, as *subsets* of $I(\Omega) \times I(\Omega)$. Clearly L is an equivalence relation, and so also is **S** given the regularity restrictions in Evans, Fraser, and Monette (1985b). Further, we can complete **C** to an equivalence relation on $I(\Omega)$ if I_1 and I_2 are viewed as being equivalent under **C** whenever there is an I_3 such that **C** gives equivalence between I_1 and I_3 and between I_2 and I_3; this was needed for the analyses in Birnbaum (1962a, 1972).

In Birnbaum's usage and in our usage to this point, the statement "principle **A** implies principle **B**" has the interpretation that "acceptance of principle **A** implies acceptance of principle **B**", which in set notation on $I(\Omega) \times I(\Omega)$ means $\mathbf{A} \supset \mathbf{B}$. By contrast we note that ordinary equivalence-relation theory would interpret "A implies B" as $A \subset B$. Thus we have the anomaly that implication among principles as used by Birnbaum, and by us to this point, is the *reverse* of the ordinary accepted usage and could be a source of confusion in discussions involving related material in mathematics and logic. Having stated this, however, we continue with the nonstandard usage, as it does conform to the manner in which these statistical principles are commonly discussed.

5. THE LIKELIHOOD PRINCIPLE

Before examining Birnbaum's proofs and our proof that **C** implies **L**, we present an example which, in its simplicity, clearly indicates the sharp discrepancies that can exist between traditional, frequency-based approaches to inference and those based on the likelihood function. This example is recorded in Fraser, Monette, and Ng (1984). Other relevant examples are found in Stein (1962), Stone (1976), and Godambe (1982).

Consider $S = \Omega = \{1, 2, \ldots\}$, and let the distribution for s be uniform on

$\{[\theta/2], 2\theta, 2\theta + 1\}$, where $[s]$ is the greatest-integer function except that $[\frac{1}{2}]$ is taken to be 1. The $S \times \Omega$ probability matrix records $p(s, \theta)$ as a function of (s, θ) on $S \times \Omega$; it is a symmetric matrix.

For a given s, the likelihood function is flat on three possible θ values $[s/2], 2s, 2s + 1$ using the preceding definition for $[\cdot]$. However, an examination of the probability matrix shows that choosing the smallest of the three possible θ-values provides a confidence procedure at level $\frac{2}{3}$; that is, one of the three θ-values (each with the same likelihood) is a 2-to-1 favourite. This can be seen by noting that for any θ the two largest s-values occur 2 times out of 3, and that the same points on $S \times \Omega$ correspond, for any s, to the smallest of the three θ-values. Thus betting on the smallest of the three possible θ-values has a 2-to-1 success probability.

It is also possible to have an inference base with this same likelihood function, but for which symmetry would show that inferences would allow no preferential choice of θ-value. For this, suppose that the original inference base had data value say $s = 20$, with possible parameter values then in $\{10, 40, 41\}$. For the new model M' let $S = \Omega = \{1, 2, \ldots\}$ and let

$$p(s, \theta) = \begin{cases} 1, & s = \theta, \\ 0 & \text{otherwise} \end{cases}$$

for θ not in $H = \{10, 11, 40, 41\}$, and

$$p(s, \theta) = \begin{cases} \frac{1}{3}, & s \in H, \quad s \neq \theta, \\ 0 & \text{otherwise} \end{cases}$$

for θ in H; and let the data point be $s = 11$. Then $\text{lik}(M, 20) = \text{lik}(M', 11)$. For this second model, however, there is full symmetry with respect to the possible parameter values, and this thus precludes preferential inferences.

Examples such as this would seem to make the likelihood principle questionable for statistical inference; for an alternative viewpoint, however, see Berger and Wolpert (1984). Also, given that **S**, **M**, and **C** are frequency-basic principles, the example makes the derivation of **L** from them seem highly paradoxical.

6. CONDITIONALITY IMPLIES LIKELIHOOD

6.1 Cross-Embedded Model

We now present a proof that **C** alone implies **L** in the context of a discrete sample space. Extensions to more general spaces can be made using results in Evans, Fraser, and Monette (1985a, 1985b).

Consider an inference base $I = (M, s^0)$ and a comparison Bernoulli inference base $I_B = (M_B, h)$ that has the same likelihood function. We build a reference model, a *cross-embedded model* M^*, using copies of the inference bases I and I_B, and as before consider a context in which both models have the same true value for θ.

For the sample space we use a matrix-type array with columns designated by the points of the sample space S of I and with rows designated by the points h, t of the Bernoulli sample space. Let p in $(0, 1)$ be any value such that

$$\frac{p}{l - p} f_\theta(s^0) \leq 1.$$

The cross-embedded model here is a minor modification of that in Evans, Fraser, and

Monette (1985a). For the first row let the probabilities be given by $pf_\theta(s)$; thus the first row has conditional model M. For the second row let the probability in cell s^0 be $1 - p - pf_\theta(s^0)$, and in some other cell, say s^1, be $pf_\theta(s^0)$:

$$\begin{array}{c|ccccc} & s^0 & s^1 & \ldots & s & \ldots \\ \hline h & pf_\theta(s^0) & pf_\theta(s^1) & \ldots & pf_\theta(s) & \ldots \\ t & 1 - p - pf_\theta(s^0) & pf_\theta(s^0) & \ldots & 0 & \ldots \end{array}$$

Note then that the conditional model for column s^0 has labels h, t and probabilities $pf_\theta(s^0)/(1 - p)$, $1 - pf_\theta(s^0)/(1 - p)$; this model is Bernoulli $(pf_\theta(s^0)/(1 - p))$, and the data value is h. By choice of p this can be made equal to a copy of the Bernoulli model M_B.

THEOREM 1. **C** *implies* **L.**

Proof. From the construction of the cross-embedded model M^* we have that the first-row indicator has probability p and is ancillary. We also have that the column-s^0 indicator is ancillary with probability $1 - p$. We are thus in the position of being able to use the conditionality principle twice:

$$\begin{aligned} \text{cont}(M^*, (h, s^0)) &= \text{cont}(M, s^0), \\ \text{cont}(M^*, (h, s^0)) &= \text{cont}(M_B, h); \end{aligned} \tag{6.1}$$

that is, $\text{cont}(M, s^0) = \text{cont}(M_B, h)$. It follows that for any two inference basis I_1 and I_2 with the same likelihood function we have that $\text{cont}(I_1) = \text{cont}(I_2)$ by transfer through the Bernoulli. Q.E.D.

6.2. On Resolution of the Paradox.

A critical step in the proof of Theorem 1 involves the use of ancillarity for the column indicator for s^0. The structure of the model M exists in the column label sample space. The indicators for column s^0 and not-(column s^0) form an ancillary statistic, and the marginal density for that ancillary is θ-free. However the general-context marginal model for that ancillary contains the essential model-M information as its understructure. In the present context, then, this (deemed) essential information is concealed in a context where the justification for conditionality says there is no information. It is by this means then that such model information is suppressed and the likelihood result obtained.

In summary we note that conditionality **C** allows model information to be concealed behind a statement that there is no relevant information.

What are the underlying mechanisms in the preceding phenomenon? First, we note that the use of the ordinary statistical model is at the core of the difficulty with Theorem 1. The cross-embedded model, treated as an ordinary model, allows the model-M information to be "absent" from the marginal model for the columns. Accordingly, it is absent at the end of the proof. Concerns with the ordinary model have been expressed at various times; see for example Fraser (1968).

Next we focus on the use of conditionality **C**. The marginal distribution for the column indicator is θ-free. The full context-based column model however would contain model-M information. Thus conditionality **C** is being used in a physical context where its justification is violated. The use of a principle when its justification is contradicted is clearly, from a general viewpoint, inappropriate.

Consider a context where a statistician who accepts **C** is presented with the inference base $(M,(h,s^0))$ of Theorem 1. In the context of the proof there are two ancillaries, one indexing the rows and the other a particular data point. We note that both of these ancillaries are maximal and thus no unique maximal ancillary exists. Accordingly these ancillaries make contradictory assertions as to which is the correct model to use for inference. It is clear that no physical context can arise where two such conflicting ancillaries each correspond to an experimental or physical ancillary. Thus the use of **C** in the theorem does not correspond to the physical justifications for **C**.

We see that the proof of the theorem proceeds precisely because of a well-known problem with **C**, namely, the lack of a unique maximal ancillary. If we were to disallow such applications, we would then not have the proof of the theorem. More formally, if we were to require that principles operate in a noncontradictory fashion, then we would not have the result in the theorem. But we would also not have the original unqualified principles.

7. ON BIRNBAUM'S PROOF THAT **C** AND **S** IMPLY **L**

We discuss briefly the proof (Birnbaum 1962a) that **C** and **S** imply **L** and the later proof (Birnbaum 1972) that **C** and **M** inply **L**.

7.1. The Mixture Model.

Consider the inference bases $I_1 = (M_1, s_1^0)$, $I_2 = (M_2, s_2^0)$, and suppose that $\text{lik}(I_1) = \text{lik}(I_2)$. Both arguments presented by Birnbaum involve a *mixture model* M^* in which with probability p_1 the model M_1 applies conditionally and with probability p_2 $(p_1 + p_2 = 1)$ the model M_2 applies conditionally.

For the use of this mixture model we note an implicit assumption not addressed in Birnbaum's discussions: that the true value of θ is the same for each of the contexts I_1 and I_2. For the mixture model M^* to have the parameter space Ω rather than $\Omega \times \Omega$ it is necessary that the two components in the mixture model have the same true value as reference. In the present development this is easily addressed: the mixture model is defined as above, with a common parameter θ for the two component conditional models, and each of these component models is viewed as a *copy* of the original models M_1 and M_2 (recall the brief discussion concerning copies in Section 2).

For the proof that **C** and **S** imply **L** we can use any nonzero values of p_1 and p_2. For the proof that **C** and **M** imply **L** we need a specific choice for the p's. We make this special choice now, thus covering both proofs. Let c be the positive constant such that $f_\theta^1(s_1^0) = cf_\theta^2(s_2^0)$ for the common likelihood points. Then let $p_1 = 1/(1+c)$ and $p_2 = c/(1+c)$.

Now consider the mixture model in more detail. We view the sample space as a two row matrix-type array with first-row values labelled by the points of S_1 and second-row values labelled by the points of S_2, and then for convenience of discussion place the value s_1^0 in the first row above the value s_2^0 in the second row. Let h, t label the rows corresponding to (say) the Bernoulli "heads" and "tails".

The mixture-model probabilities for the first row are given by $p_1 f_\theta^1(s_1)$; thus the first row has conditional model M_1. Similarly the probabilities for the second row are given by $p_2 f_\theta^2(s_2)$; thus the second row has conditional model M_2. The marginal model for the row label h, t is of course Bernoulli (p_1).

Note that the rows are ancillary. Thus by conditionality principle **C** we have that

$$\begin{aligned} \text{cont}(I_1) &= \text{cont}(M_1, s_1^0) = \text{cont}(M^*, s_1^0), \\ \text{cont}(I_2) &= \text{cont}(M,_2, s_2^0) = \text{cont}(M^*, s_2^0). \end{aligned} \tag{7.1}$$

The linking of the right sides is then provided by the principles **S** or **M**.

The points s_1^0 and s_2^0 have the same likelihood function and thus belong to a contour of a sufficient statistic; it follows that the right sides of (7.1) are equal by **S**, and thus $\text{cont}(I_1) = \text{cont}(I_2)$. Alternatively, we note that the points s_1^0 and s_2^0 have the same probability function in $M^* : p_1 f_\theta^1(s_1^0) = p_2 f_\theta^2(s_2^0)$ by our construction. It follows then that the right sides of (7.1) are equal by **M**, and thus that $\text{cont}(M_1) = \text{cont}(M_2)$.

Various reservations have been raised previously concerning these results. Fraser (1963) noted that when we consider the class of models that are transformation models, then two inference bases that have the same likelihoods can give quite different inferences when the transformation structure is taken into account. Of course, Birnbaum's discussion does not include these models, and thus this does not invalidate the results.

Durbin (1970) noted that Birnbaum's results do not hold if we require that the inference process proceed by making the reductions first via sufficiency to the minimal sufficient statistic and then via conditionality. Then in the first proof we would not be able to assert (6.1) after (6.2). On the other hand, there seems to be no compelling reason to prefer this order of application, and thus we cannot view this as resolving the paradoxical result.

Kalbfleisch (1975, p. 255) notes that if the inference process proceeded by first conditioning on an experimental ancillary, the paradoxical result would be avoided. Some suggestions in this direction may be found in Basu (1964) and in an example discussed in Fraser (1973). The experimental ancillary is treated as providing an instance of necessary reduction in Fraser (1979, §3.2).

One could view the direction indicated by the experimental ancillary as providing the fundamental resolution of the issues and paradoxes raised by Birnbaum's arguments. With the class $\mathcal{I}(\Omega)$, however, the paradoxical result is not voided. For clearly we see that the notion of an experimental ancillary, or structural ancillary from a different viewpoint, requires specifications beyond that provided by $\mathcal{I}(\Omega)$, that is, it requires the specification of a certain variable or certain subspace as having distinguished properties. A related step in this direction is provided by the additional elements of the structural models, beyond those in the ordinary model.

7.2. The Difficulties in the Arguments.

A critical step in the two arguments occurs when s_1^0 and s_2^0 are taken to have the same inference implications in the mixture model, in our case by **S** and in the other case by **M**. The points, however, clearly label the corresponding component models M_1 and M_2.

Thus we note that sufficiency **S** and mathematical equivalence **M** allow model information to be concealed behind a statement that there is no relevant information. This is the essential mechanism by which **L** is derived from **C** and **S** and from **C** and **M**.

Alternatively, consider a context where a statistician who accepts **C** and **S** is presented with the mixture inference base $(M^*, (h, s_1^0))$. Conditionality indicates that the relevant model for inference about θ is given by M_1. On the other hand, application of **S** establishes—as is clearly seen via the sufficient statistic $t(x, s) = \{(h, s_1^0)\}$ if $(x, s) \in \{(h, s_1^0), (t, s_2^0)\}$ and $t(x, s) = (x, s)$ otherwise—that the information as to which model has occurred is irrelevant information for inferences about θ. The statistician is presented with contradictory recommendations from these principles. Clearly the principles interact in such a context in a way that leads to doubts as to the validity of their application and accordingly doubts as to the validity of the proof. Recalling the justifications of **S** and **C**, we note that the random system used for postrandomization in S and prerandomization in C can be taken to be one and the same in the applications of these principles in the proofs.

It is not hard to see that such a conflict between **C** and **S** occurs when and only when two points in a sample space give rise to the same likelihood function and yet give different values for some ancillary variable. This is what happens in Birnbaum's proof (1962a), and if such applications are disallowed, then the proof fails. It is also apparent that such a conflict exists between **M** and **C** in the Birnbaum (1972) proof.

From Section 6 we recall that conditionality **C** in Theorem 1 allowed model information to be concealed behind a statement that there is no relevant information, and this was the mechanism by which **L** was derived from **C** in Theorem 1. Further, **C** was used in a context which does not correspond to its justification.

The three proofs thus have essentially the same mechanism by which information beyond the likelihood function is eliminated. And furthermore, in each case the usage of the particular principle runs counter to the motivation used to support the principle.

8. ALTERNATIVES

What has gone wrong? Clearly, the use of the ordinary statistical model (2.1) restricts what information can be carried by a model, and in the context of the mixture model actually eliminates the component-model identification. Thus the general deficiency of the ordinary statistical model provides the mechanism for the proofs giving the paradoxical results.

8.1. More than an Equivalence Relation.

Given the disturbing consequences of Birnbaum's formulation of the common principles, we examine more closely the meaning and uses of a principle. We recall that $\text{cont}(I_1) = \text{cont}(I_2)$ means that I_1 and I_2 contain the same information concerning the parameter θ. We first question to what degree a statistical principle is merely the statement of an equivalence.

Consider the sufficiency principle. Given $I = (M, s)$ and a statistic t sufficient for M, the sufficiency principle as described above asserts that (M, s) and $(M', t(s))$ contain the same information. Operationally, however, the principle **S** seems to imply more: that we should replace (M, s) by $(M', t(s))$ for purposes of inference. For associated with any inference base is a wealth of inference procedures that can commonly be invoked, and in replacing (M, s) by $(M', t(s))$ we are restricting this class, unless of course t is trivial. In this sense *sufficiency* can be viewed as an operation step towards cont, and would be more than a mere statement of equivalence.

A further point concerns the justifications for a principle, which, after all, motivate our using it. Necessarily, in any given context where we are going to use the principle, we must be sure that we are using it in a way that corresponds to its justification. Otherwise the support for the principle is being compromised. It is partly for this reason that we introduce below two modifications for sufficiency and conditionality. A further point arises when more than one principle is used: that the justifications for the principles be not in conflict; for this would mean yet again that the support for the principles is compromised.

Birnbaum did not address these aspects of the principles, only treating them as equivalence relations. Accordingly his proofs, and also the present proof that **C** implies **L**, allow the use of the principles in contexts where the justification for the principles is violated.

Such applications are clearly inappropriate and indicate at least that some clarification is needed of the principle, or of the application context $\mathcal{I}(\Omega)$, or of both.

8.2. *Modified Principles.*

In an attempt to address the issues raised by the proofs, we introduce two variations on the sufficiency principle:

S(as): As the conditional model given a sufficient statistic contains no information concerning the parameter, cont(I_1) = cont(I_2) if I_2 can be obtained from I_1 via a sufficiency map.

S(if): If the conditional model given a sufficient statistic contains no information concerning the parameter, cont(I_1) = cont(I_2) if I_2 can be obtained from I_1 via a sufficiency map.

For Birnbaum's proof (1962a) we have noted that sufficiency **S** is used in a context where the justification for the principle is violated. If we use the modification S(as), we have that this modified principle is not true. If we use the modification S(if), we have that the modified principle is not applicable. These provide two possible resolutions of the paradox.

For Birnbaum's proof (1972) we have noted that mathematical equivalence **M** is used in a context where two points in fact are *not* mathematically equivalent, that is, where component-model identification is admitted over and above the ordinary model representation of the mixture model.

For our current Theorem 1, we propose two modifications of **C** pàralleling those introduced for sufficiency. For this let **C**(as) and **C**(if) be the principle **C** preceded by "as the marginal model of an ancillary contains no information concerning the parameter" and "if the marginal model of the ancillary contains no information concerning the parameter", respectively.

Suppose now for Theorem 1 that we replace **C** by one of its modifications. We note that the principle **C**(as) is then seen to be false. On the other hand, for the principle **C**(if) we see that the principle is not applicable. However, this modified principle **C**(if) may be so qualified as to be not easily or usefully checked.

9. CONCLUDING COMMENTS

We have examined Birnbaum's arguments from **C**, **S**, and **M** to **L** and presented a proof from **C** to **L**. The results are paradoxical, and we have probed the sources of these disturbing results.

From one viewpoint the restriction to an ordinary space-algebra-measure model provides the means by which information beyond the likelihood function can be made to disappear. Some preliminary discussion concerning what is an appropriate statistical model may be found in Fraser (1979); the present result reemphasizes these issues.

From a general viewpoint we can state that conditionality—as intended by its justifications—does not imply likelihood. The direction proposed by Kalbfleisch (1975), following the lead of earlier authors, provides a key to reformulating conditionality in the context of more general models.

At a minimum we have shown that Birnbaum's use of **S** and **M** and our use of **C** in Theorem 1 are contrary to the intentions of the principles, as judged by the relevant supporting and motivating examples. From this viewpoint we can state that the intentions of **S** and **C** do not imply **L**. This points to a more qualified approach to the basic principles.

REFERENCES

Barnard, G., and Godambe, V.P. (1982). Allan Birnbaum 1923–1976. *Amer. Statist.*, 10, 1033–1039.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā*, 15, 377–380.

Basu, D. (1958). On statistics independent of a sufficient statistic. *Sankhyā*, 20, 223–226.

Basu, D. (1964). Recovery of ancillary information. *Sankhyā*, 26A, 3–16.

Basu, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā Ser. A*, 37, 1–71.

Berger, J.O., and Wolpert, R.L. (1984). *The Likelihood Principle. Lecture Notes—Monograph Series, Vol 6. Inst. Math. Statist.*

Birnbaum, A. (1962a). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.*, 57, 269–332.

Birnbaum, A. (1962b). Another view on the foundations of statistics. *Amer. Statist.*, 16, 17–21.

Birnbaum, A. (1969). Concepts of statistical evidence. *Philosophy, Science and Method* (S. Morgenbesser, P. Suppes, and M. White, *eds.*). 112–143.

Birnbaum, A. (1970). On Durbin's modified principle of conditionality. *J. Amer. Statist. Assoc.*, 65, 402–403.

Birnbaum, A. (1972). More on concepts of statistical evidence. *J. Amer. Statist. Assoc.*, 67, 858–886.

Brenner, D.; Fraser, D.A.S., and Monette, G. (1981). Theories of inference or simple additives. *Statist. Hefte*, 22, 231–234.

Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–372.

Dawid, A.P. (1977). Conformity of inference patterns. *Recent Developments in Statistics* (J.R. Barra, et al., *eds.*). North-Holland, Amsterdam, 244–256.

Durbin, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Assoc.*, 65, 395–398.

Evans, M.; Fraser, D.A.S., and Monette, G. (1985a). Mixtures, embeddings and ancillarity. *Canad. J. Statist.*, 13, 1–6.

Evans, M.; Fraser, D.A.S., and Monette, G. (1985b). Regularity conditions for statistical models. *Canad. J. Statist.*, 13, 137–144.

Fraser, D.A.S. (1963). On the sufficiency and likelihood principles. *J. Amer. Statist. Assoc.*, 58, 641–647.

Fraser, D.A.S. (1968). A black box or a comprehensive model. *Technometrics*, 10, 219–229.

Fraser, D.A.S. (1973). The elusive ancillary. *Multivariate Statistical inference* (D.G. Kabe and R.P. Gupta, *eds.*). North-Holland, Amsterdam, 41–48.

Fraser, D.A.S. (1976). *Probability and Statistics, Theory and Applications.* DAI, Univ. of Toronto Textbook Store, Toronto.

Fraser, D.A.S. (1979). *Inference and Linear Models.* DAI, Univ. of Toronto Textbook Store, Toronto.

Fraser, D.A.S. (1984). Structural models. *Encyclopedia of Statistical Science. Volume 6.* (N.L. Johnson and S. Kotz, *eds.*). Wiley, New York.

Fraser, D.A.S.; Monette, G., and Ng, K.-W. (1984). Marginalization, likelihood and structural models. *Multivariate Analysis VI* (P.R. Krishnaiah, *ed.*). North-Holland, Amsterdam.

Godambe, V.P. (1979). On Birnbaum's axion of mathematically equivalent experiments. *J. Roy. Statist. Soc. Ser. B*, 41, 789–827.

Godambe, V.P. (1982). Ancillarity principle and a statistical paradox. *J. Amer. Statist. Assoc.*, 77, 931–933.

Hajek, J. (1967). On basic concepts of statistics. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, 139–162.

Hartigan, J.A. (1967). The likelihood and invariance principles. *Ann. Math. Statist.*, 3, 533–539.

Joshi, V.M. (1976). A note on Birnbaum's theory of the likelihood principle. *J. Amer. Statist. Assoc.*, 71, 345–346.

Kalbfleisch, J.D. (1975). Sufficiency and conditionality. *Biometrika*, 62, 251–259.

Pratt, J.W. (1961). Book review. *J. Amer. Statist. Assoc.*, 56, 163–166.

Rudin, W. (1974). *Real and Complex Analysis.* McGraw-Hill, New York.

Savage, L.J. (1970). Comments on a weakened principle of conditionality. *J. Amer. Statist. Assoc.*, 65, 399–401.

Stein, C. (1962). A remark on the likelihood principle. *J. Amer. Statist. Soc.*, A125, 565–568.

Stone, M. (1976). Strong inconsistencies from uniform priors (with discussion). *J. Amer. Statist. Assoc.*, 71, 114–125.

Received 9 November 1984
Revised 21 January 1986
Accepted 31 January 1986

Department of Statistics
100 St. George Street
University of Toronto
Toronto, Ontario M5S 1A1

Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario N2L 3G1

Department of Mathematics
York University
4700 Keele Street
Downsview, Ontario M3J 1P3

DISCUSSION

J.D. KALBFLEISCH, *University of Waterloo*

I would like to congratulate the authors on a stimulating and insightful paper. The careful analysis and discussion throws much light on the general issues involved, and I find myself largely in agreement with their views and conclusions.

I found the example in Section 5 to be compelling at first, but decreasingly so as I considered variations on the theme. I question the summary statement that, for given s, $\theta = [s/2]$ is a 2:1 favourite. Consider an analogous example but with parameter space truncated at $2N + 1$. It is still the case that $P(\theta = [s/2]) \geqq 2/3$ for all θ, but the likelihood is concentrated on $\theta = [s/2]$ if $s > N$ and identical to that obtained in Section 5 if $s \leqq N$. One should clearly be unwilling to assert, in this case, that $\theta = [s/2]$ is a 2:1 favourite once s is observed; if $s > N$ the odds are infinite in its favour, and if $s \leqq N$ the odds would presumably be less than 2:1 (in fact, 1:2 would appear more descriptive). For $s \leqq N$ and N large, this example is very similar to the infinite case in Section 5, yet we seem to be led to different inferences. The identifiable subsets for which the probability statement should differ are clear in the finite case; are there similar considerations in the infinite case which are somehow missed?

I do have reservations about the appropriateness of L and agree with the authors that the restrictive and abstract model M lies at the root of the many derivations of L from frequentist principles. All these derivations have a common mechanism of proof involving the introduction of artificial mixtures and the application of principles to these mixtures in an automatic way without, as the authors indicate, regard to their motivations. This arbitrariness in the sample space and in experiment definition makes the emergence of a principle like L unsurprising. An essential additive to M, in my view, is the scientific context of the experiment. This includes many aspects: for example, its purpose, particular decision processes which led to the experiment and random elements in the experiment design. This was my motivation in formulating an experimental conditionality principle with the aim of defining the experiment actually performed (see Kalbfleisch 1975). The formulation is incomplete — perhaps necessarily so — and ambiguities may still exist, but