Electronic Journal of Statistics
Vol. 7 (2013) 2645–2655
ISSN: 1935-7524
DOI: 10.1214/13-EJS857

What does the proof of Birnbaum's theorem prove?

Michael Evans

Department of Statistics
University of Toronto
e-mail: mevans@utstat.utoronto.ca
url: www.utstat.utoronto.ca/mikevans

Abstract: Birnbaum's theorem, that the sufficiency and conditionality principles entail the likelihood principle, has engendered a great deal of controversy and discussion since the publication of the result in 1962. In particular, many have raised doubts as to the validity of this result. Typically these doubts are concerned with the validity of the principles of sufficiency and conditionality as expressed by Birnbaum. Technically it would seem, however, that the proof itself is sound. In this paper we use set theory to formalize the context in which the result is proved and show that in fact Birnbaum's theorem is incorrectly stated as a key hypothesis is left out of the statement. When this hypothesis is added, we see that sufficiency is irrelevant, and that the result is dependent on a well-known flaw in conditionality that renders the result almost vacuous.

AMS 2000 subject classifications: Primary 62A01; secondary 62F99.
Keywords and phrases: Sufficiency, conditionality, likelihood, relations, equivalence relations.

Received February 2013.

1. Introduction

A result presented in [2], and referred to as Birnbaum's theorem, is very well-known in statistics. This result says that a statistician who accepts both the sufficiency S and conditionality C principles must also accept the likelihood principle L and conversely. The result has always been controversial primarily because it implies that a frequentist is forced to ignore the repeated sampling properties of any inferential procedures they use. Given that both S and C seem quite natural to many frequentist statisticians while L does not, the result is highly paradoxical.

Various concerns have been raised about the proof of the result. For example, in [4] Durbin argued that the theorem fails to hold whenever C is restricted by requiring that any ancillaries used must be functions of a minimal sufficient statistic. In [11] it is argued that C should only be applicable when the value of the ancillary statistic used to condition is actually a part of the experimental make-up. This is called the weak conditionality principle. In [5] it is argued that Birnbaum's theorem, and a similar result that accepting C alone is equivalent to accepting L, are invalid because the specific uses of S and C in proving these results can be seen to be based on flaws in their formulations. For example, Birnbaum's theorem requires a use of S and C where the information discarded by S as irrelevant, which is the primary motivation for S, is exactly the information used by C to condition on and so identifies the discarded information as highly relevant. As such S and C contradict each other. We note that this is precisely what Durbin's restriction on the ancillaries avoids. Furthermore, the result established in [5] that C alone is equivalent to L, depends on the lack of a unique maximal ancillary which can be seen as an essential flaw in C. Also, see [9, 1] and [8] for various concerns about the formulation of the theorem. The validity of the theorem and the principles are discussed in [13]. It is argued in [12] that, in the context of a repeated sampling formulation for statistics, we cannot simultaneously have S and C true, as when S is true then C is false and when C is true then S is false. A proof that avoids some of the objections raised by others is provided in [6].

Many of these reservations are essentially with the hypotheses to the theorem and suggest that Birnbaum's theorem should be rejected because the hypotheses are either not acceptable or have been misapplied. It is the purpose of this paper to provide a careful set-theoretic formulation of the context of the theorem. When this is done we see that there is a hypothesis that needs to be formally acknowledged as part of the statement of Birnbaum's theorem. With this addition, the force of the result is lost and the paradox disappears. The same conclusions apply to the result that C is equivalent to L and, in fact, this is really the only result as S is redundant in the statement of Birnbaum's theorem when the additional hypothesis is formally acknowledged.

For our discussion it is important that we stick as closely as possible to Birnbaum's formulation. To discuss the proof, however, we have to make certain aspects of Birnbaum's argument mathematically precise that are somewhat vague in his paper. It is always possible then that someone will argue that we have done this in a way that is not true to Birnbaum's intention. We note, however, that this is accomplished in a very simple and direct way. If there is another precise formulation that makes the theorem true, then it is necessary for a critic of how we do this to provide that alternative.

A basic step missing in [2] was to formulate the principles as relations on the set $\mathcal{I}$ of all model and data combinations. So $\mathcal{I}$ is the set of all *inference bases* $I = (E, x)$ where $E = (\mathcal{X}_E, \{f_{E,\theta} : \theta \in \Theta_E\}), \mathcal{X}_E$ is a sample space, $\{f_{E,\theta} : \theta \in \Theta_E\}$ is a collection of probability density functions on $\mathcal{X}_E$, with respect to some support measure μ_E on $\mathcal{X}_E$, indexed by $\theta \in \Theta_E$, and $x \in \mathcal{X}_E$ is the observed data. We will ignore all measure-theoretic considerations as they are not essential for any of the arguments. If the reader is concerned by this, then we note that the collection of models where $\mathcal{X}_E$ and Θ_E are finite and μ_E is counting measure is rich enough to produce the paradoxical result. So we can restrict our discussion here to the case where $\mathcal{X}_E$ and Θ_E are finite. In spite of our restrictions, most of our development applies in more general circumstances although it is fair to acknowledge that the discussion in Birnbaum, and here, is restricted to parametric statistical inference.

We note that expressing the principles as relations was part of [5] but this is taken further here. In Section 2 we discuss the meaning and use of relations generally and relate this to the statistical application. In Section 3 we apply our discussion of relations to Birnbaum's theorem. In Section 4 we draw some conclusions.

2. Relations

In [2] Birnbaum considered the principles S, C and L as equivalence relations on the set $\mathcal{I}$ of all inference bases $I = (E, x)$ as described in Section 1. Birnbaum's theorem can be interpreted mathematically as a statement about relationships existing among these equivalence relations but the paper is not careful about the usage of this terminology relying instead on intuition. So in this section we give a precise definition of what is meant by relations and equivalence relations, give some basic examples and discuss a key concept necessary for a precise statement of Birnbaum's theorem, namely, the smallest equivalence relation containing a relation. We use [7] as the source of our basic definitions.

Suppose D is a set. A *relation* R with domain D is a subset $R \subset D \times D$. Saying $(x, y) \in R$ means that the objects x and y have a property in common.

Example 1. Suppose D is the set of students enrolled at a specific university at a specific point in time. Let R_1 be defined by $(x, y) \in R_1$ when x and y are enrolled in the same program. Let R_2 be defined by $(x, y) \in R_2$ whenever x and y are enrolled in, or have been enrolled, in the same program.

A relation R is *reflexive* if $(x, x) \in R$ for all $x \in D$, *symmetric* if $(x, y) \in R$ implies $(y, x) \in R$, and *transitive* if $(x, y) \in R, (y, z) \in R$ implies that $(x, z) \in R$. If a relation R is reflexive, symmetric and transitive, then R is called an *equivalence relation*. Note that an equivalence relation R is not empty, unless D is empty, since $\{(x, x) : x \in D\} \subset R$. In Example 1, clearly R_1 is an equivalence relation and, while R_2 is reflexive and symmetric, it is not typically transitive, as students change programs, and so may not be an equivalence relation. While $(x, y) \in R$ implies that x and y are related, perhaps by the possession of some property, when R is an equivalence relation this implies that x and y possess the property to the same degree. We say that relation R on D *implies* relation R' on D whenever $R \subset R'$. In Example 1, we have that $R_1 \subset R_2$.

In the context of Birnbaum's theorem we take $D = \mathcal{I}$ and consider a *statistical relation* as a relation $R \subset \mathcal{I} \times \mathcal{I}$ and call such an R a *statistical principle* when R is an equivalence relation. Of course, not all such R are meaningful statistical relations or principles as they may simply satisfy the mathematical properties without possessing any statistical content. For example, select any two inference bases I_1 and I_2 with the same parameter space, then formally $R = \{(I_1, I_1), (I_2, I_2), (I_1, I_2), (I_2, I_1)\}$ is a statistical principle but clearly not for any particular statistical reason. In general, it would seem that statistical relations need to be defined, as in [2], based on some concept associated with statistical evidence. So, if $(I_1, I_2) \in R$, then I_1 and I_2 are supposed to contain equivalent amounts of statistical information about the unknown distribution or, equivalently, the unknown true value of the model parameter. As with

Birnbaum, we make no attempt to give a precise definition of what statistical information means.

If R is a relation on D, then the equivalence relation $\bar{R}$ generated by R is the smallest equivalence relation containing R. We see that $\bar{R}$ is the intersection of all equivalence relations on D containing R. We will use the following characterization of $\bar{R}$.

Lemma 1. *If R is a reflexive relation on D, then*

$$\begin{aligned}\bar{R} = \{(x,y) : \exists\, n, x_1, \ldots, x_n \in D \textit{ with } x = x_1, y = x_n \textit{ and } \\ (x_i, x_{i+1}) \in R \textit{ or } (x_{i+1}, x_i) \in R\}.\end{aligned}$$

Proof. Since $R \subset \bar{R}$ we have that $\bar{R}$ is reflexive. If $(x,y) \in \bar{R}$, then there exists $n, x_1, \ldots, x_n \in D$ with $x = x_1, y = x_n$ and $(x_i, x_{i+1}) \in R$ or $(x_{i+1}, x_i) \in R$ and so $(y,x) \in \bar{R}$ using $x_i' = x_{n-i+1}$ instead of the x_i. If $(x,y), (y,z) \in \bar{R}$, then we have $(x,z) \in \bar{R}$ simply by concatenating the chains that put $(x,y) \in \bar{R}$ and $(y,z) \in \bar{R}$. □

Note that for statistical relations R, as characterizations of statistical information, it makes sense to assume that R is reflexive since inference base I must contain the same amount of information as itself. Lemma 1 is used to prove that certain statistical relations are not equivalence relations.

It may be that $\bar{R}$ does not have a meaningful interpretation, at least as it relates to the property being expressed by R. In Example 1, $\bar{R}_2$ is difficult to interpret and surely goes beyond the idea that R_2 is perhaps trying to express, namely, that two students have some common interests. In fact, it is entirely possible that $\bar{R}_2 = D \times D$ and so says nothing. As another example, suppose that $D = \{2, 3, 4, \ldots\}$ and $(x,y) \in R$ when x and y have a common factor bigger than 1. Then R is reflexive and symmetric but not transitive. If $x, y \in D$ then $(x, xy) \in R, (xy, y) \in R$ so $\bar{R} = D \times D$ and $\bar{R}$ is saying nothing. It seems that each situation, where we extend a relation R to an equivalence relation, must be examined to see whether or not this extension has any meaningful content for the application. As we discuss in Section 3, this process of extending a relation to be an equivalence relation is implicit in Birnbaum's result and in the result of [5], and as such, suggests that these results have no substantive inferential content.

In Section 3 we need to consider the union $R_1 \cup R_2$ of relations R_1 and R_2 on D. In general, the union of equivalence relations is not an equivalence relation. We have the following result.

Lemma 2. $\overline{\bar{R}_1 \cup \bar{R}_2} = \overline{R_1 \cup R_2}$.

Proof. We have that $R_1 \cup R_2 \subset \bar{R}_1 \cup \bar{R}_2$ so $\overline{R_1 \cup R_2} \subset \overline{\bar{R}_1 \cup \bar{R}_2}$ while $\bar{R}_1 \subset \overline{R_1 \cup R_2}, \bar{R}_2 \subset \overline{R_1 \cup R_2}$ implies $\overline{\bar{R}_1 \cup \bar{R}_2} \subset \overline{R_1 \cup R_2}$. □

This says that the equivalence relation generated by the union of relations is equal to the equivalence relation generated by the union of the corresponding generated equivalence relations.

3. Statistical relations and principles

We now consider several statistical relations and the statistical principles generated by them. Birnbaum's theorem is a statement about the relations among them.

The *likelihood relation* L on $\mathcal{I}$ is defined by $(I_1, I_2) \in L$ whenever $\Theta_{E_1} = \Theta_{E_2}$ and there exists $c > 0$ such that $f_{E_1,\theta}(x_1) = cf_{E_2,\theta}(x_2)$ for every $\theta \in \Theta_{E_1}$. We have the following obvious result.

Lemma 3. *L is a statistical principle.*

The likelihood principle as stated can be generalized in an obvious way. For we may have $I_1 = (E_1, x_1)$ and $I_2 = (E_2, x_2)$, a bijection $h : \Theta_{E_1} \to \Theta_{E_2}$ and a constant $c > 0$, such that $f_{E_1,\theta}(x_1) = cf_{E_2,h(\theta)}(x_2)$ for every $\theta \in \Theta_{E_1}$. It then seems reasonable to consider $(I_1, I_2) \in L$. We will ignore this generalization here as it is not relevant to our arguments. Effectively we will require that I_1 and I_2 have the same parameter space anytime we consider them to be related via a statistical relation.

We consider the definition of the *sufficiency relation* S. First we show that a minimal sufficient statistic always exists for the models discussed here, namely, $\mathcal{X}_E$ and Θ_E are finite. Furthermore, we suppose that for each $x \in \mathcal{X}_E$ there exists $\theta \in \Theta_E$ such that $f_{E,\theta}(x) > 0$, so we don't allow any points in $\mathcal{X}_E$ that can't be observed. We say that two points $x_1, x_2 \in \mathcal{X}_E$ are equivalent whenever there exists constant $k > 0$ such that $f_{E,\theta}(x_1) = kf_{E,\theta}(x_2)$ for every $\theta \in \Theta$ and denote the equivalence class containing x by $[x]$. We have the following result.

Lemma 4. *$T(x) = [x]$ is a minimal sufficient statistic for E.*

Proof. If $z \in [x]$, then there exists $k(z) > 0$ such that $f_{E,\theta}(z) = k(z)f_{E,\theta}(x)$ for every θ. Suppose that θ is true. If $f_{E,\theta}(x) > 0$, then $f_{E,\theta}(z) > 0$ for every $z \in [x]$ and the conditional probability of x given $T(x) = [x]$ based on $f_{E,\theta}$ equals

$$f_{E,\theta}(x)/\sum_{z\in[x]} f_{E,\theta}(z) = f_{E,\theta}(x)/\sum_{z\in[x]} k(z)f_{E,\theta}(x) = 1/\sum_{z\in[x]} k(z). \tag{1}$$

If $f_{E,\theta}(x) = 0$, then $f_{E,\theta}(z) = 0$ for every $z \in [x]$ and so the probability of $[x]$ based on $f_{E,\theta}$ is 0. Therefore, we can define the conditional probability of x given $T(x) = [x]$ based on $f_{E,\theta}$ arbitrarily. There is a θ' such that $f_{E,\theta'}(x) > 0$ and the conditional probability of x given $T(x) = [x]$ based on $f_{E,\theta'}$ equals (1). So if we define the conditional probability of x given $T(x)$ based on $f_{E,\theta}$ by (1), this conditional probability is independent of θ and we have that T is sufficient.

Now suppose that U is a sufficient statistic for E. Then

$$f_{E,\theta}(x) = f_E(x \,|\, U(x))f_{E,\theta,U}(U(x)) \tag{2}$$

where $f_E(\cdot \,|\, U(x))$ is the conditional probability function given $U(x)$ and $f_{E,\theta,U}$ is the marginal for U. Since $f_{E,\theta}(x) > 0$ for at least one θ, we must have that

$f_E(x \,|\, U(x)) > 0$. If $U(x_1) = U(x_2)$, then $f_{E,\theta,U}(U(x_1)) = f_{E,\theta,U}(U(x_2))$ and from (2)

$$f_{E,\theta}(x_1) = f_E(x_1 \,|\, U(x_1)) f_{E,\theta,U}(U(x_2)) = \frac{f_E(x_1 \,|\, U(x_1))}{f_E(x_2 \,|\, U(x_2))} f_{E,\theta}(x_2),$$

which implies that $T(x_1) = T(x_2)$. This proves that T is a minimal sufficient statistic. □

Any 1-1 function of a minimal sufficient statistic is also minimal sufficient. So we can always take our minimal sufficient statistic to be real-valued here since $[x]$ takes only finitely many values. Also, if $T$and T' are both minimal sufficient for model E, then $T = h \circ T'$ for some function h, since T' is sufficient and T is minimal sufficient. If $T'(x_1) \neq T'(x_2)$ but $h(T'(x_1)) = h(T'(x_2))$, then T' would not be minimal sufficient and so h must be 1-1.

Let T_i denote a minimal sufficient statistic for model E_i with marginal model E_{i,T_i}. Define the sufficiency relation by $(I_1, I_2) \in S$ whenever there is a 1-1 map h between the sample spaces of E_{1,T_1} and E_{2,T_2} such that $E_{1,T_1} = E_{2,h\circ T_2}$ and $T_1(x_1) = h(T_2(x_2))$. We have the following result.

Lemma 5. *S is a statistical principle and $S \subset L$.*

Proof. Consider inference base $I = (E, x)$ and suppose T, T' are minimal sufficient statistics for E. Then, as we have discussed, there is a 1-1 function h such that $T = h \circ T'$ which implies $E_T = E_{h\circ T'}$ and $T(x) = h(T'(x))$ so $(I, I) \in S$ and S is reflexive. If $(I_1, I_2) \in S$ via h, then $E_{1,T_1} = E_{2,h\circ T_2}$ and so T_1 and $h \circ T_2$ have the same distribution for each θ. This implies that $h^{-1} \circ T_1$ and T_2 have the same distribution for each θ and so $E_{1,h^{-1}\circ T_1} = E_{2,T_2}$. Also, $T_1(x_1) = h(T_2(x_2))$ implies $T_2(x_2) = h^{-1}(T_1(x_1))$ so $(I_2, I_1) \in S$ which proves S is symmetric. If $(I_1, I_2) \in S$ via h_1 and $(I_2, I_3) \in S$ via h_2, then $(I_1, I_3) \in S$ via $h = h_1 \circ h_2$, and so S is transitive.

Now (2) implies that a likelihood function obtained from (E, x) is proportional to a likelihood function obtained from $(E_T, T(x))$ when T is a minimal sufficient statistic for E. When $(I_1, I_2) \in S$, then $E_{1,T_1} = E_{2,h\circ T_2}$ and $T_1(x_1) = h(T_2(x_2))$ which implies that a likelihood function obtained from $(E_{1,T_1}, T_1(x_1))$ is proportional to a likelihood function obtained from $(E_{2,T_2}, T_2(x_2))$. Therefore, a likelihood function obtained from I_1 is proportional to a likelihood function obtained from I_2 so $(I_1, I_2) \in L$. We conclude that $S \subset L$. □

A statistic a for model E is *ancillary* if the marginal model induced by a is given by one probability distribution, namely, the distribution of a is independent of $\theta \in \Theta_E$. For $x \in \mathcal{X}_E$ the *conditional model* given $a(x)$ is $\{f_{E,\theta}(\cdot \,|\, a(x)) : \theta \in \Theta_E\}$ where $f_{E,\theta}(\cdot \,|\, a(x))$ is the density for the data given $a(x)$. The *conditionality relation* C is defined by $(I_1, I_2) \in C$ whenever $\Theta_{E_1} = \Theta_{E_2}$, $x_1 = x_2$ and there exists ancillary statistic a for E_1 such that the conditional model given $a(x_1)$ is E_2, or with roles of I_1 and I_2 reversed. Basically the conditionality relation is saying that whether we use the conditional model given an ancillary or the unconditional model we should have the same inferences. There are

Table 1
Unconditional distributions

(x_1, x_2)	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
$f_{E,1}(x_1, x_2)$	$1/6$	$1/6$	$2/6$	$2/6$
$f_{E,2}(x_1, x_2)$	$1/12$	$3/12$	$5/12$	$3/12$

Table 2
Conditional distributions given $U = 1$

(x_1, x_2)	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
$f_{E,1}(x_1, x_2 \,\vert\, U = 1)$	$1/2$	$1/2$	0	0
$f_{E,2}(x_1, x_2 \,\vert\, U = 1)$	$1/4$	$3/4$	0	0

Table 3
Conditional distributions given $V = 1$

(x_1, x_2)	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
$f_{E,1}(x_1, x_2 \,\vert\, V = 1)$	$1/3$	0	$2/3$	0
$f_{E,2}(x_1, x_2 \,\vert\, V = 1)$	$1/6$	0	$5/6$	0

numerous examples where it is apparent that the conditional model is more appropriate for assessing the uncertainties associated with inferences, see [3]. We have the following result.

Lemma 6. *C is reflexive and symmetric but is not transitive and $C \subset L$.*

Proof. The reflexivity, symmetry and $C \subset L$ are obvious. The lack of transitivity follows via a simple example. Consider the model E with $\mathcal{X}_E = \{1, 2\}^2, \Theta_E = \{1, 2\}$ and with $f_{E,\theta}$ given by Table 1. Now note that $U(x_1, x_2) = x_1$ and $V(x_1, x_2) = x_2$ are both ancillary and the conditional models, when we observe $(x_1, x_2) = (1, 1)$, are given by Tables 2 and 3. The only ancillary for both these conditional models is the trivial ancillary (the constant map). Therefore, there are no applications of C that lead to the inference base I_2, given by Table 2 with data $(1, 1)$, being related to the inference base I_3, given by Table 3 with data $(1, 1)$. But both of I_2 and I_3 are related under C to the inference base I_1 given by Table 1 with data $(1, 1)$. This establishes the result. □

Note that even under relabellings, the inferences bases I_2 and I_3 in Lemma 6 are not equivalent.

If we are going to say that $(I_1, I_2) \in C$ means that I_1 and I_2 contain an equivalent amount of information under C, then we are forced to expand C to $\bar{C}$ so that it is an equivalence relation. But this implies that the two inference bases I_2 and I_3 presented in the proof of Lemma 6 contain an equivalent amount of information and yet they are not directly related via C. Rather they are related only because they are conditional models obtained from a supermodel that has two essentially different maximal ancillaries. An ancillary a is maximal if, whenever $a = g \circ a'$ and a' is ancillary, then g is a bijection.

Saying that such models contain an equivalent amount of statistical information is clearly a substantial generalization of C. Note that, for the example in the proof of Lemma 6, when $(1, 1)$ is observed, the MLE is $\hat{\theta}(1, 1) = 1$. To measure the accuracy of this estimate we can compute the conditional probabilities

Table 4
The model E_1^*

	x_1	x_{10}	x_{100}	$\cdots$
$i=1$	$pf_{E_1,\theta}(x_1)$	$pf_{E_1,\theta}(x_{10})$	$pf_{E_1,\theta}(x_{100})$	$\cdots$
$i=0$	$1-p-pf_{E_1,\theta}(x_1)$	$pf_{E_1,\theta}(x_1)$	0	$\cdots$

based on the two inference bases, namely,

$$P_1(\hat{\theta}(x_1,x_2)=1\,|\,U=1)=1/2,\ P_2(\hat{\theta}(x_1,x_2)=2\,|\,U=1)=3/4$$
$$P_1(\hat{\theta}(x_1,x_2)=1\,|\,V=1)=1/3,\ P_2(\hat{\theta}(x_1,x_2)=2\,|\,V=1)=5/6$$

and so the accuracy of $\hat{\theta}$ is quite different depending on whether we use I_2 or I_3. It seems unlikely that we would interpret these inference bases as containing an equivalent amount of information in a frequentist formulation of statistics. As noted in Section 2, there is no reason why we have to accept the equivalences given by a generated equivalence relation unless we are certain that this equivalence relation expresses the essence of the basic relation. It seems clear that there is a problem with the assertion that $(I_1,I_2)\in\bar{C}$ means that I_1 and I_2 contain an equivalent amount of information without further justification.

We now follow a development similar to that found in [5] to prove the following result.

Theorem 7. *$C\subset\bar{C}=L$ where the first containment is proper.*

Proof. Clearly $C\subset\bar{C}$ and this containment is proper by Lemma 6. If $(I_1,I_2)\in\bar{C}$, then Lemma 1 implies $(I_1,I_2)\in L$ since $C\subset L$ and so $\bar{C}\subset L$. Now suppose that $(I_1,I_2)\in L$. We have that $f_{E_1,\theta}(x_1)=cf_{E_2,\theta}(x_2)$ for every θ for some $c>0$. Assume first that $c>1$. Now construct a new inference base $I_1^*=(E_1^*,(1,x_1))$ where $\mathcal{X}_{E_1^*}=\{0,1\}\times\mathcal{X}_{E_1}$, and $\{f_{E_1^*,\theta}:\theta\in\Theta_{E_1}\}$ is given by Table 4 where $x_{10},x_{100},\ldots$ are the elements of $\mathcal{X}_{E_1}$ not equal to x_1 and $p\in[0,1)$ satisfies $p/(1-p)=1/c$.

Then we see that $U(i,x)=i$ is ancillary as is V given by $V(i,x)=1$ when $x=x_1$ and $V(i,x)=0$ otherwise. Conditioning on $U(i,x)=1$ gives that $(I_1^*,I_1)\in C$ while conditioning on $V(i,x)=1$ gives that $(I_1^*,I)\in C$ where $I=((\{0,1\},\{p_\theta:\theta\in\Theta_{E_1}\}),1)$ and p_θ is the Bernoulli$(f_{E_1,\theta}(x_1)/c)$ probability function. Now, using I_2 we construct I_2^* by replacing p by $1/2$ and $f_{E_1,\theta}(x_1)$ by $f_{E_2,\theta}(x_2)$ in Table 4 and obtain that $(I_2^*,I)\in C$ since $f_{E_1,\theta}(x_1)/c=f_{E_2,\theta}(x_2)$. Using Lemma 1 we have that $(I_1,I_2)\in\bar{C}$. If $c\le 1$ we start the construction process with I_2 instead. This proves that $\bar{C}=L$. □

The proof that $L\subset\bar{C}$ relies on discreteness. This was weakened in [5] and even further weakened in [10].

We now show that Birnbaum's proof actually establishes the following result.

Theorem 8. *$S\cup C\subset L\subset\overline{S\cup C}$.*

Proof. The first containment is obvious. For the second suppose that $(I_1,I_2)\in L$. We construct a new inference base $I=(E,y)$ from I_1 and I_2 as follows. Let E

be given by $\mathcal{X}_E = (1, \mathcal{X}_{E_1}) \cup (2, \mathcal{X}_{E_2})$,

$$f_{E,\theta}(1,x) = \begin{cases} (1/2)f_{E_1,\theta}(x) & \text{when } x \in \mathcal{X}_{E_1} \\ 0 & \text{otherwise,} \end{cases}$$
$$f_{E,\theta}(2,x) = \begin{cases} (1/2)f_{E_2,\theta}(x) & \text{when } x \in \mathcal{X}_{E_2} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$T(i,x) = \begin{cases} (i,x) & \text{when } x \notin \{x_1,x_2\} \\ \{x_1,x_2\} & \text{otherwise} \end{cases}$$

is sufficient for E and so $((E,(1,x_1)),(E,(2,x_2))) \in S$ by Lemma 5. Also, $h(i,x) = i$ is ancillary for E and thus

$$\begin{aligned} ((E,(1,x_1)),(E_1,x_1)) &\in C, \\ ((E,(2,x_2)),(E_2,x_2)) &\in C. \end{aligned}$$

Then by Lemma 1 we have that $((E_1,x_1),(E_2,x_2)) \in \overline{S \cup C}$ and we are done. □

Note that Birnbaum's proof only proves the containments with no equalities but we have the following result.

Theorem 9. *$S \cup C$ is properly contained in L while $L = \overline{S \cup C}$.*

Proof. We show that $S \cup C \subset L$ is proper. Suppose that E_1 has $\mathcal{X}_{E_1} = \{0,1\}$, $\Theta_{E_1} = \{1/5, 1/3\}$ with $f_{E_1,\theta}(x) = \theta^x(1-\theta)^{1-x}$ and E_2 has $\mathcal{X}_{E_1} = \{0,1,2\}$, $\Theta_{E_2} = \{1/5, 1/3\}$ with $f_{E_2,\theta}(0) = \theta$, $f_{E_2,\theta}(1) = \theta(1-\theta)$ and $f_{E_2,\theta}(2) = (1-\theta)^2$. Suppose further that we observe $x_1 = 1$ and $x_2 = 0$ so $f_{E_1,\theta}(1) = \theta = f_{E_2,\theta}(0)$. Note that the full data is minimal sufficient for both E_1 and E_2 and that both of these models have only trivial ancillaries. Therefore, if $I_1 = (E_1, 1)$ and $I_2 = (E_2, 0)$, we have that $(I_1, I_2) \notin S$, $(I_1, I_2) \notin C$ but $(I_1, I_2) \in L$ which proves that $S \cup C$ is properly contained in L.

To prove that the second containment is exact we have, using Lemma 1, that $(I_1, I_2) \in \overline{S \cup C}$ implies that I_1 and I_2 give rise to proportional likelihoods as this is true for each element of $S \cup C$ and so $\overline{S \cup C} \subset L$. □

So we do not have, as usually stated for Birnbaum's Theorem, that S and C are together equivalent to L, but we do have that $\overline{S \cup C}$ is equivalent to L. Acceptance of $\overline{S \cup C}$ is not entailed, however, by acceptance of both S and C as we have to examine the additional relationships added to $S \cup C$ to see if they make sense. If one wishes to say that acceptance of S and C implies the acceptance of $\overline{S \cup C}$, then a compelling argument is required for these additions and this seems unlikely.

From Theorems 7 and 8 we have the following Corollary.

Corollary 10. *$S \cup C \subset \bar{C} = L$ where the first containment is proper. Furthermore, $S \subset \bar{C}$ and this containment is proper.*

A direct proof that $S \subset \bar{C}$ has been derived in [10]. It is interesting to note that Corollary 10 shows that the existence of S in the modified statement of Birnbaum's theorem, where we require that we accept all the equivalences generated by S and C, is irrelevant. This is a reassuring result as it is unlikely that S is defective but it is almost certain that C is defective, at least as currently stated.

As with the proof of Birnbaum's Theorem, the proof that $C = L$ provided in [5] is really a proof that $\bar{C} = L$. This can be seen from the proof of Theorem 7. So accepting the relation C is not really equivalent to accepting L unless we agree that the additional elements of $\bar{C}$ make sense. This is essentially equivalent to saying that it doesn't matter which maximal ancillary we condition on and it is unlikely that this is acceptable to most frequentist statisticians. This is illustrated by the discussion concerning the example in Lemma 6.

As noted in [4], requiring that any ancillaries used in an application of C be functions of a minimal sufficient statistic voids Birnabum's proof, as the ancillary statistic used in the proof of Theorem 7 is not a function of the sufficient statistic used in the proof. It is not clear, however, what this restriction does to the result $\bar{C} = L$, but we note that there are situations where there exist nonunique maximal ancillaries which are functions of the minimal sufficient statistic. In these circumstances we would still be forced to conclude the equivalence of inference bases derived by conditioning on the different maximal ancillaries if we reasoned as in [5]. Of course, we are arguing here that the result requires the statement of an additional hypothesis.

4. Conclusions

We have shown that the proof in [2] did not prove that S and C lead to L. Rather the proof establishes that $\overline{S \cup C} = L$ and this is something quite different. The statement of Birnbaum's theorem in prose should have been: if we accept the relation S and we accept the relation C *and* we accept all the equivalences generated by S and C together, then this is equivalent to accepting L. The essential flaw in Birnbaum's theorem lies in excluding this last hypothesis from the statement of the theorem. The same qualification applies to the result proved in [5] where the statement of the theorem should have been: if we accept the relation C *and* we accept all the equivalences generated by C, then this is equivalent to accepting L.

The way out of the difficulties posed by Birnbaum's theorem, and the result relating C and L, is to acknowledge that additional hypotheses are required for the results to hold. Certainly these results seem to lose their impact when they are correctly stated and we realize that an equivalence relation generated by a relation is not necessarily meaningful. It is necessary to provide an argument as to why the generated equivalence relation captures the essence of the relation that generates it and it is not at all clear how to do this in these cases.

As we have noted, the essential result in all of this is $\bar{C} = L$ and this has some content albeit somewhat minor. Furthermore, the proof of this result is based on a defect in C, namely, it is not an equivalence relation due to the general nonexistence of unique maximal ancillaries. As such it is hard to accept

C as stated as any kind of characterization of statistical evidence. Given the intuitive appeal of this relation in some simple examples, however, resolving the difficulties with C still poses a major challenge for a frequentist theory of statistics.

Acknowledgements

The author thanks the reviewers for a number of constructive comments.

References

[1] Barndorff-Nielsen, O. E. (1995) Diversity of evidence and Birnbaum's theorem (with discussion). *Scand. J. Statist.*, 22(4), 513–522. MR1363227

[2] Birnbaum, A. (1962) On the foundations of statistical inference (with discussion). *J. Amer. Stat. Assoc.*, 57, 269–332. MR0138176

[3] Cox, D. R. and Hinkley, D. V. (1974) Theoretical Statistics. Chapman and Hall. MR0370837

[4] Durbin, J. (1970) On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Stat. Assoc.*, 654, 395–398.

[5] Evans, M., Fraser, D. A. S. and Monette, G. (1986) On principles and arguments to likelihood (with discussion). *Canad. J. of Statistics*, 14, 3, 181–199. MR0859631

[6] Gandenberger, G. (2012) A new proof of the likelihood principle. To appear in the *British Journal for the Philosophy of Science*.

[7] Halmos, P. (1960) Naive Set Theory. Van Nostrand Reinhold Co. MR0114756

[8] Helland, I. S. (1995) Simple counterexamples against the conditionality principle. *Amer. Statist.*, 49, 4, 351–356. MR1368487

[9] Holm, S. (1985) Implication and equivalence among statistical inference rules. In *Contributions to Probability and Statistics in Honour of Gunnar Blom*. Univ. Lund, Lund, 143–155. MR0795054

[10] Jang, G. H. (2011) The conditionality principle implies the sufficiency principle. Working paper.

[11] Kalbfleisch, J. D. (1975) Sufficiency and conditionality. *Biometrika*, 62, 251–259. MR0386075

[12] Mayo, D. (2010) An error in the argument from conditionality and sufficiency to the likelihood principle. In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science* (D. Mayo and A. Spanos eds.). Cambridge University Press, Cambridge, 305–314. MR2640508

[13] Robins, J. and Wasserman, L. (2000) Conditioning, likelihood, and coherence: A review of some foundational concepts. *J. Amer. Stat. Assoc.*, 95, 452, 1340–1346. MR1825290