Birnbaum's Theorem Redux

Sérgio Wechsler, Carlos A. de B. Pereira and Paulo C. Marques F.

Instituto de Matemática e Estatística – Universidade de São Paulo – Brasil
sw@ime.usp.br, cpereira@ime.usp.br, pmarques@ime.usp.br

Abstract. We revisit Birnbaum's results on the Likelihood Principle, reorganizing and making a few formal changes which address some of the criticisms at the original development. The meaning of the results for different theories of statistical inference is exemplified, and the role of the Likelihood Principle as means to assess the consistency of those theories is emphasized.

Keywords: Likelihood Principle, Statistical Inference.
PACS: 02.50.-r, 02.50.Tt.

This paper presents the results of Birnbaum [1, 2, 3, 4, 5] on the Likelihood Principle. As for the chosen inferential principles and the logical equivalences among them, we largely follow Basu [3]. The main change is basically stylistic: we give up on the controversial use of a vaguely defined evidence functional, and establish the needed equivalences among realizations of experiments passing to the quotient of the appropriate class by means of an equivalence relation.

The idea is to construct a formal structure that allows us to represent the inferential statements of any possible statistical method of parametric inference. Our first step is to define *experiment* and derived concepts. By an experiment we mean a statistical model with its standard ingredients: the sample space $\mathscr{X}$, some suitable σ-field $\mathscr{B}$ of subsets of $\mathscr{X}$, the parametric space Θ, and a family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability measures on $(\mathscr{X}, \mathscr{B})$. We consider only *discrete* models, for which $\mathbb{P}_\theta$ has countable support, for every θ in Θ. We also abbreviate the notation of the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ by $\{f_\theta\}$, where each f_θ is a probability function on $\mathscr{X}$.

DEFINITION. An *experiment* E is a triple $(\mathscr{X}, \Theta, \{f_\theta\})$. The class of all such triples will be denoted by $\mathscr{E}$. For some fixed Θ, the class of all *experiments about* Θ is defined as $\mathscr{E}_\Theta := \{E \in \mathscr{E} : c_2 E = \Theta\}$, where $c_i : \mathscr{S}_1 \times \cdots \times \mathscr{S}_n \to \mathscr{S}_i : (S_1, \ldots, S_n) \mapsto S_i$, for $i = 1, \ldots, n$.

Thus, the proverbial "quantity of interest" assumes values on a given Θ, and we want to compare realizations of experiments in $\mathscr{E}_\Theta$.

DEFINITION. A *realization* of an experiment E is a pair (E, x), where x is some point of $c_1 E$. The class of all *realizations of experiments about* Θ is $\mathscr{R}_\Theta := \bigcup_{E \in \mathscr{E}_\Theta} (\{E\} \times c_1 E)$.

EXAMPLE 1. Suppose that we will observe the results of trials in a clinical experiment. One way to perform the experiment is to fix some $n > 0$ and observe the number of successes in n trials. This corresponds to the formal experiment $E_0 = (\mathscr{X}_0, \Theta, \{f^0_\theta\})$, where $\mathscr{X}_0 = \{0, 1, 2, \ldots, n\}$, $\Theta = (0, 1)$, and $f^0_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, for each x in $\mathscr{X}_0$. Another possibility is to fix some $r > 0$, run the trials until we get r failures, and observe the number of successes. In this case, we have the experiment $E_1 = (\mathscr{X}_1, \Theta, \{f^1_\theta\})$, where

$\mathscr{X}_1 = \{0,1,2,\ldots\}$, $\Theta = (0,1)$, and $f_\theta^1(x) = \binom{r+x-1}{r-1}\theta^x(1-\theta)^r$, for each x in $\mathscr{X}_1$. If x_0 and x_1 are points of $\mathscr{X}_0$ and $\mathscr{X}_1$, respectively, we may want to compare, for different methods of inference, the inferential content of the realizations (E_0,x_0) and (E_1,x_1). □

Given some method of inference, to compare the inferential content of realizations in $\mathscr{R}_\Theta$, we introduce an equivalence relation $\simeq$ on $\mathscr{R}_\Theta$. By definition, the only constraints to $\simeq$ are that it must be

- *Reflexive*: $(E,x) \simeq (E,x)$;
- *Symmetric*: if $(E,x) \simeq (E',x')$, then $(E',x') \simeq (E,x)$; and
- *Transitive*: if $(E,x) \simeq (E',x')$ and $(E',x') \simeq (E'',x'')$, then $(E,x) \simeq (E'',x'')$,

where (E,x), (E',x'), and (E'',x'') are realizations in $\mathscr{R}_\Theta$. Apart from that, $\simeq$ can be any equivalence relation on $\mathscr{R}_\Theta$. Let $E = (\mathscr{X},\Theta,\{f_\theta\})$ and $E' = (\mathscr{X}',\Theta,\{f'_\theta\})$ be two experiments. If $R = (E,x)$ and $R' = (E',x')$ are realizations in $\mathscr{R}_\Theta$, the following are examples of the type of equivalence relations in which we will be interested.

EXAMPLE 2. Say that $R \simeq_1 R'$ if and only if $\arg\max_\theta f_\theta(x) = \arg\max_\theta f'_\theta(x')$. This relation $\simeq_1$ corresponds to the inferential statement of maximum likelihood estimation. □

EXAMPLE 3. For a suitable σ-field $\mathscr{A}$ of subsets of Θ, suppose that we have a prior density ξ with respect to a measure λ on $(\Theta,\mathscr{A})$, and that we will evaluate the posterior densities with respect to the measure λ for both realizations using Bayes Theorem: $\pi(\theta \mid x) \propto f_\theta(x)\xi(\theta)$, and $\pi'(\theta \mid x') \propto f'_\theta(x')\xi(\theta)$, a.s. $[\lambda]$. Say that $R \simeq_2 R'$ if and only if $\pi(\cdot \mid x) = \pi'(\cdot \mid x')$ a.s. $[\lambda]$. This relation $\simeq_2$ captures the idea that in Bayesian inference the posterior is the inferential statement *par excellence*. Another possibility is to say that $R \simeq_3 R'$ if and only if $\int_\Theta \theta \cdot \pi(\theta \mid x)d\lambda(\theta) = \int_\Theta \theta \cdot \pi'(\theta \mid x')d\lambda(\theta)$, that is, if and only if the Bayes estimates with quadratic losses are the same for both realizations. Of course, $\simeq_2 \subset \simeq_3$. □

In his original paper [1], Birnbaum used an evidence functional to establish this sort of equivalence. Objections to the vague nature of this functional were raised, as in the comments of LeCam in [4]. We think that an equivalence relation is the right set-theoretical tool here, as Birnbaum himself recognized in a later paper [2].

DEFINITION. Consider an experiment $E = (\mathscr{X},\Theta,\{f_\theta\})$, and let $T : \mathscr{X} \to \mathscr{Y}$ be a bijective transformation. The *experiment E transformed by T* is defined as the triple $TE := (\mathscr{Y},\Theta,\{g_\theta\})$, where $g_\theta(y) = f_\theta(T^{-1}y)$, for each y in $\mathscr{Y}$, and every θ in Θ.

The following Invariance Principle (I) of Basu [3] simply states that the particular labeling of the sample space should be immaterial to the inferential statements.

PRINCIPLE. (Invariance) Let E be an experiment in $\mathscr{E}_\Theta$, and x a value in c_1E. If we have any bijective transformation $T : \mathscr{X} \to \mathscr{Y}$, then $(E,x) \simeq (TE,Tx)$.

LEMMA. Let E be an experiment in $\mathscr{E}_\Theta$, and let x_0 and x_1 be two values in c_1E. Assume that (I) holds. If $f_\theta(x_0) = f_\theta(x_1)$, for every θ in Θ, then $(E,x_0) \simeq (E,x_1)$.

PROOF. Consider the mapping

$$T : \mathscr{X} \to \mathscr{X} : x \mapsto \begin{cases} x_1 & , \text{ if } x = x_0 \ , \\ x_0 & , \text{ if } x = x_1 \ , \\ x & , \text{ otherwise .} \end{cases}$$

T is bijective. The experiment E transformed by T is the triple $TE = (\mathscr{X}, \Theta, \{g_\theta\})$, where $g_\theta(x) = f_\theta(T^{-1}x)$, for every x in $\mathscr{X}$, and each θ in Θ. Since, by hypothesis, $f_\theta(x_0) = f_\theta(x_1)$, for every θ in Θ, we have that $g_\theta = f_\theta$, and therefore $\{g_\theta\} = \{f_\theta\}$. In other words, $E = (\mathscr{X}, \Theta, \{f_\theta\})$ and $TE = (\mathscr{X}, \Theta, \{g_\theta\})$ are the same triple, and this implies that $(TE, Tx_0) = (E, x_1)$, because $Tx_0 = x_1$, by the definition of T. In the other hand, (I) says that $(E, x_0) \simeq (TE, Tx_0)$, and then we conclude that $(E, x_0) \simeq (E, x_1)$. □

The former Lemma was stated as the Principle of Mathematical Equivalence by Birnbaum in [2].

DEFINITION. Let $E_0 = (\mathscr{X}_0, \Theta, \{f^0_\theta\})$ and $E_1 = (\mathscr{X}_1, \Theta, \{f^1_\theta\})$ be two experiments, let α be a real constant in $[0,1]$, and let J be a random variable with Bernoulli distribution of parameter α. The *α-mixed experiment of E_0 and E_1* is the triple $E_\alpha := (\mathscr{X}_\alpha, \Theta, \{f^\alpha_\theta\})$, where $\mathscr{X}_\alpha = (\{0\} \times \mathscr{X}_0) \cup (\{1\} \times \mathscr{X}_1)$, and we have $f^\alpha_\theta(j, x_j) = \alpha^j (1-\alpha)^{1-j} f^j_\theta(x_j)$, for $j = 0, 1$.

Birnbaum introduced the following Conditionality Principle (C).

PRINCIPLE. (Conditionality) Let $E_0 = (\mathscr{X}_0, \Theta, \{f^0_\theta\})$ and $E_1 = (\mathscr{X}_1, \Theta, \{f^1_\theta\})$ be two experiments, and let x_0 and x_1 be two values in $\mathscr{X}_0$ and $\mathscr{X}_1$, respectively. For every α in $[0,1]$, if E_α is the α-mixed experiment of E_0 and E_1, then $(E_\alpha, (j, x_j)) \simeq (E_j, x_j)$, for $j = 0, 1$.

The idea contained in (C) is that if we use any coin to decide which of two experiments will be taken to effect, the inferential statements of this mixed experiment must be exactly the same as those of the experiment actually performed, without any special reference to the other, unperformed, experiment. What matters is what actually happened. That (C) should be a natural requirement for a Bayesian is clear from the fact that it is archetypical in Bayesian inference to condition only on the observed facts. It is our opinion that any reasonable theory of inference should satisfy (C).

We finally come to Birnbaum's statement of the Likelihood Principle.

PRINCIPLE. (Likelihood) Let $E_0 = (\mathscr{X}_0, \Theta, \{f^0_\theta\})$ and $E_1 = (\mathscr{X}_1, \Theta, \{f^1_\theta\})$ be two experiments. If there are two values x_0 and x_1 in $\mathscr{X}_0$ and $\mathscr{X}_1$, respectively, and there is a real number $k = k(x_0, x_1) > 0$, such that $f^0_\theta(x_0) = k \cdot f^1_\theta(x_1)$, for every θ in Θ, then $(E_0, x_0) \simeq (E_1, x_1)$.

EXAMPLE 4. Consider the realizations of the experiments in Example 1, with $n = 12$, $x_0 = 9$, $r = 3$, and $x_1 = 9$. Suppose that we want to test the hypothesis $H_0 : \theta \leq 1/2$ against the alternative $H_1 : \theta > 1/2$, and that, for any two realizations R and R' in $\mathscr{R}_\Theta$, we define $R \simeq_4 R'$ if and only if the classical p-values of both realizations are exactly the same. For the realization (E_0, x_0) we have the p-value $p_0 = \sum_{x=9}^{12} \binom{12}{x} \left(\frac{1}{2}\right)^{12} \approx 0.0729$. For the realization (E_1, x_1) the p-value is $p_1 = \sum_{x=9}^{\infty} \binom{2+x}{2} \left(\frac{1}{2}\right)^{x+3} \approx 0.0327$. We see that both realizations are not equivalent with respect to $\simeq_4$. In this setting, we may also define $R \simeq_5 R'$ if and only if for both realizations we decide in the same way for rejection or non-rejection of H_0, after adopting some standard significance level. Now, with the traditional $0,05$ significance level, we would not reject H_0 in realization (E_0, x_0), but we would reject H_0 in realization (E_1, x_1). All of this is at odds with (L), since both realizations have likelihoods proportional to $\theta^9(1-\theta)^3$. Notice that if a Bayesian has

a prior $Beta(a,b)$ for the parameter, both realizations would give the same posterior $Beta(9+a,3+b)$, and so $(E_0,x_0) \simeq_2 (E_1,x_1)$. □

Imagine a statistician who feels that satisfaction of (I) and (C) is a reasonable requirement for his theories of inference, but considers (L) to be of a different nature, seeing no good reason to abide by it. The following Theorem is Basu's version [3] of Birnbaum's remarkable discovery that our fellow statistician would sooner or later face a contradiction.

THEOREM. (C) and (I) if and only if (L).

PROOF. ($\Rightarrow$) Let $E_0 = (\mathscr{X}_0, \Theta, \{f^0_\theta\})$ and $E_1 = (\mathscr{X}_1, \Theta, \{f^1_\theta\})$ be two experiments, and let x_0 and x_1 be two values in $\mathscr{X}_0$ and $\mathscr{X}_1$, respectively. Suppose that (C) and (I) hold, and that, by hypothesis, there is a real number $k > 0$ such that $f^0_\theta(x_0) = k \cdot f^1_\theta(x_1)$, for every θ in Θ. Take $\alpha = k/(k+1)$. Consider E_α, the α-mixed experiment of E_0 and E_1. From the definition of E_α, we have that, for every θ in Θ,

$$f^\alpha_\theta(0,x_0) = \frac{1}{k+1} \cdot f^0_\theta(x_0) = \frac{k}{k+1} \cdot f^1_\theta(x_1) = f^\alpha_\theta(1,x_1)\,.$$

Therefore, it follows from the proved Lemma that $(E_\alpha,(0,x_0)) \simeq (E_\alpha,(1,x_1))$. From (C), and the transitivity of $\simeq$, it follows that $(E_0,x_0) \simeq (E_1,x_1)$, and then we conclude that (L) holds.

($\Leftarrow$) Now, we will show that (L) entails (C). By the definition of E_α, we know that $f^\alpha_\theta(j,x_j) = \alpha^j(1-\alpha)^{1-j} f^j_\theta(x_j)$, for $j = 0,1$. That is, the realizations $(E_\alpha,(0,x_0))$ and (E_0,x_0) have proportional likelihoods, with $k = 1-\alpha$, and the same happens for the realizations $(E_\alpha,(1,x_1))$ and (E_1,x_1), with $k = \alpha$. It follows from (L) that $(E_\alpha,(j,x_j)) \simeq (E_j,x_j)$, for $j = 0,1$, and therefore (C) holds. To show that (L) entails (I), let $E = (\mathscr{X},\Theta,\{f_\theta\})$ be an experiment, let x be a value in $\mathscr{X}$, and let $T : \mathscr{X} \to \mathscr{Y}$ be any bijective mapping. Denote by X the random vector whose values we observe in experiment E. In the transformed experiment $TE = (\mathscr{Y},\Theta,\{g_\theta\})$ we observe the values of the random vector $Y = T(X)$. Since T is bijective, $X = x$ if and only if $Y = Tx$, and therefore $\{X = x\} = \{Y = Tx\}$. Then, $f_\theta(x) = P_\theta\{X = x\} = P_\theta\{Y = Tx\} = g_\theta(Tx)$, for every θ in Θ. Hence, the likelihoods of the realizations (E,x) and (TE,Tx) are the same, and it follows from (L), with $k = 1$, that $(E,x) \simeq (TE,Tx)$, showing that (I) holds. □

As we have seen in Example 4, classical methods of inference may violate (L). We see this as a side-effect of the recurrent attempt in classical statistics to equate the concept of probability with that of *hypothetical* frequencies. This makes classical statistics highly counterfactual, with its inference statements depending heavily on the nature of sample space points which were not observed, in such a way that consistency with (C), and so with (L), is in general impossible. It should be noticed that Bayesian inference, with its subjective definition of probability, establishes consistency between personal beliefs and *actually observed* frequencies through results such as De Finetti's version of the Law of Large Numbers [6].

In our opinion, the intuitive meaning of (C) and (I), and the logical equivalence established by the Theorem above, shows that satisfaction of (L) is a necessary condition for any cogent method of statistical inference.

ACKNOWLEDGMENTS

We warmly thank Professor Luis Gustavo Esteves for his detailed revision of the manuscript, and many stimulating discussions on these issues over the years. We also thank Professor Júlio Michael Stern for asking us to submit this paper to these Proceedings, and Professor John Skilling for his helpful suggestions.

REFERENCES

1. A. Birnbaum, "On the Foundations of Statistical Inference", with discussion, *Journal of the American Statistical Association* **57**, 269–326 (1962).
2. A. Birnbaum, "More on Concepts of Statistical Evidence", *Journal of the American Statistical Association* **67**, 858–861 (1972).
3. D. Basu, "Statistical Information and Likelihood", with discussion, *Sankhya A* **37**, 1–71 (1975).
4. J. O. Berger and R. Wolpert, "The Likelihood Principle" (Second Edition), Institute of Mathematical Statistics Lecture Notes – Monograph Series (Hayward, CA, IMS, 1988).
5. A. P. Dawid, "Conformity of Inference Patterns", in *Recent Developments in Statistics*, edited by J. R. Barra et al. (North-Holland, Amsterdam, 1977).
6. M. J. Schervish, *Theory of Statistics* (Springer, New York, 1995).