RONALD N. GIERE*

ALLAN BIRNBAUM'S CONCEPTION OF STATISTICAL EVIDENCE

Allan Birnbaum died in London in the early summer of 1976. He was 53. Although he was by training and profession a statistician, his intellectual interests were more philosophical than mathematical. During the past fifteen years most of his efforts were devoted to the foundations of statistics. His paper on the Neyman-Pearson Theory (Birnbaum [41]) is only the latest, though now unfortunately the last, of several papers developing a point of view that Birnbaum thought to represent both the theory and practice of the majority of reflective theoretical statisticians. In the following pages I will outline the development of Birnbaum's views on statistical inference. I hope this will help those unfamiliar with the earlier papers better to understand this last paper and also provide additional motivation for looking at the earlier papers as well.

Birnbaum was unusual among statisticians in that he actively sought intellectual contact with philosophers as well as with methodologists in various sciences. Thus several of his papers, especially the later ones, were written so as to require relatively little technical expertise in statistics. They must be read very carefully, however, for they are written in a style that is sometimes complex and usually understates the significance of the point being made. The style is an accurate reflection of the man.

1. SUFFICIENCY, CONDITIONALITY AND LIKELIHOOD

In December of 1961 Birnbaum presented the paper 'On the Foundations of Statistical Inference' (Birnbaum [19]) at a special discussion meeting of the American Statistical Association. Among the discussants was L. J. Savage who pronounced it "a landmark in statistics". Explicitly denying any "intent to speak with exaggeration or rhetorically", Savage described the occasion as "momentous in the history of statistics". "It would be hard", he said, "to point to even a handful of comparable

Synthese **36** (1977) 5–13.

events" (Birnbaum [19], pp. 307–8). The reasons for Savage's enthusiasm are obvious. Birnbaum claimed to have shown that two principles widely held by non-Bayesian statisticians (sufficiency and conditionality) jointly imply an important consequence of Bayesian statistics (likelihood). The outlines of Birnbaum's analysis are easily sketched.

The basic concept is that of the 'evidential meaning', $\mathrm{Ev}(E, x)$, of an outcome, x, relative to an experiment, E. The experiment is itself a complex entity consisting of a parameter space, Ω, a sample space, S, and a set of distribution functions, $f(x, \theta)$, where $x \in S$ and $\theta \in \Omega$. For example, tossing a bent coin 50 times to determine the probability of heads is an experiment. Ω is the set of values $0 \leqslant p \leqslant 1$; S is the set of possible numbers of heads, i.e., $\{0, 1, \ldots, 50\}$; f is the corresponding set of binomial distribution functions. Thus $\mathrm{Ev}(E, 40)$ would be the 'evidential meaning' of getting 40 heads in this experiment.

The concept of 'evidential meaning' is explicated in terms of axioms on the relation $\mathrm{Ev}(E, x)$. Indeed, the basic axioms are *invariance* conditions; they state when two evidential meanings are the same. The first one is sufficiency.

The Principle of Sufficiency (S): If E is a specified experiment with outcomes S; if $t = t(x)$ is any sufficient statistic; and if E' is the experiment, derived from E, in which any outcome of E is represented only by the corresponding value $t = t(x)$; then for each x, $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', t(x))$ (Birnbaum [19], p. 270).

A difficulty with this principle is that it assumes there is a clear definition of a sufficient statistic. The idea behind a sufficient statistic is that $t(x)$ contain all the 'relevant information' in x. But then one needs a measure of relevant information. Birnbaum was aware of these problems and in later papers developes notions of sufficiency independent of the concept of a sufficient statistic. He was right, however, in thinking that a great many statisticians would subscribe to a principle of "the irrelevance of observations independent of a sufficient statistic".

To get at the idea of conditionality, imagine two possible experiments that differ only in the size of the sample chosen. Suppose the experiment actually to be performed is determined by the outcome of a stochastic

process, e.g., the toss of a coin. The whole sequence may be viewed as a 'mixture' of experiments in which E_1 is performed with probability p and E_2 with probability $1-p$. The principle of conditionality says that once one of the two experiments is selected, the evidential meaning of the outcome is just as if the other possibility had never existed. More formally:

The Principle of Conditionality (C): If E is any experiment having the form of a mixture of component experiments E_h, then for each outcome (E_h, x_h) of E we have $\mathrm{Ev}(E, (E_h, x_h)) = \mathrm{Ev}(E_h, x_h)$. (Birnbaum [19], p. 271).

Informally, this principle asserts "the irrelevance of (component) experiments not actually performed". Birnbaum cites Fisher and D. R. Cox as being among the many statisticians who have endorsed such a principle.

Finally, consider the function $f(x, \theta)$ as a function of θ for fixed x. The resulting function is called the likelihood function for θ given x. This function is not a probability function, though for fixed θ and x it has the same value as the corresponding probability function. The likelihood principle asserts the equivalence in evidential meaning of two experiments whose likelihood functions differ at most by a constant multiple. More formally:

The Likelihood Principle (L): If E and E' are any two experiments with the same parameter space, represented respectively by density functions $f(x, \theta)$ and $g(y, \theta)$; and if x and y are any respective outcomes determining the same likelihood function; then $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', y)$ (Birnbaum [19], p. 271).

Informally, the likelihood principle asserts "the irrelevance of outcomes not actually observed". Stated positively, the likelihood principle asserts that the whole evidential meaning of the observed outcome is given by the likelihood of the outcome that in fact occurs, and no other features of the experiment are relevant.

Except for Bayesians, not many statisticians have endorsed the likelihood principle. Barnard, and perhaps Fisher, are the most notable advocates. But many statisticians have endorsed both sufficiency and

conditionality. Thus the impact of Birnbaum's result that (S) and (C) hold if and only if (L) holds. Actually, (C) implies (S), so (C) and (L) are equivalent. It seems that one cannot have (C) without getting (L) as well.

The connection with Bayesian inference is immediate. According to Bayes' Theorem, the posterior probability of h given x is proportional to the prior probability of h times the likelihood of h relative to x. Thus the observed outcome of the experiment enters only by means of its likelihood function, which is what the likelihood principle asserts. Of course, Bayesian inference goes beyond the likelihood principle, but Savage was certainly right in regarding endorsement of the likelihood principle as a big step in the direction of Bayesian inference. Birnbaum, however, repeatedly refused to take what Savage and others regarded as the obvious next step. Indeed, he seems almost immediately to have backed away from his qualified acceptance of the likelihood principle as an "appropriate characterization of statistical evidence" (Birnbaum, [19], p. 323).

2. REJECTION OF THE LIKELIHOOD PRINCIPLE

Birnbaum's reputation among statisticians is based mostly on the 1962 paper. Thus most regard him as the author of an original attempt to 'justify' the likelihood principle. This is a reasonable reading of the 1962 paper. It is clear from unpublished manuscripts, however, that as early as 1964 Birnbaum did not regard the likelihood principle as providing an adequate interpretation of the concept of statistical evidence. This is fairly clear in his article 'Likelihood' in the *International Encyclopedia of the Social Sciences* (Birnbaum [26]). It is absolutely explicit in the letter he published in *Nature* in response to A. W. F. Edwards' short exposition of the likelihood approach in that journal. Here Birnbaum wrote: I am not now among the 'modern exponents' of the likelihood concept.[1] He then went on to say:

> If there has been "one rock in a shifting scene" of general statistical thinking and practice in recent decades, it has not been the likelihood concept, as Edwards suggests, but rather the concept by which confidence limits and hypothesis tests are usually interpreted, which we may call the confidence concept of statistical evidence. This concept is not part of the Neyman-Pearson theory of tests and confidence region estimation, which denies any role to concepts of statistical evidence, as Neyman consistently insists. The confidence concept

takes from the Neyman-Pearson approach techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data. (The absence of a comparable property in the likelihood and Bayesian approaches is widely regarded as a decisive inadequacy.) The confidence concept also incorporates important but limited aspects of the likelihood concept: the sufficiency concept, expressed in the general refusal to use randomized tests and confidence limits when they are recommended by the Neyman-Pearson approach; and some applications of the conditionality concept. It is remarkable that this concept, an incompletely formalized synthesis of ingredients borrowed from mutually incompatible theoretical approaches, is evidently useful continuously in much critically informed statistical thinking and practice (Birnbaum [29]).

So much for the likelihood principle. But what of the confidence concept?

3. THE CONFIDENCE CONCEPT

The 1962 paper contains no reference to a 'confidence concept' of statistical evidence. There is no confidence principle. There is, however, a discussion of 'intrinsic confidence methods', a reinterpretation of standard methods of testing and interval estimation in terms of likelihood ratios. These ideas had been developed by Birnbaum in earlier papers. So the idea that something like 'confidence' should be important was in the background all along. In an unpublished 1964 manuscript entitled 'The Anomalous Concept of Statistical Evidence', Birnbaum introduced the following principle:

Unbiasedness criterion for a mode of evidential interpretations (U): Systematically misleading or inappropriate interpretations shall be impossible; that is, under no θ shall there be high probability of outcomes interpreted as 'strong evidence against θ'.

This criterion expresses the Neyman-Pearson interpretation of Fisher's significance level which was later generalized to include error probabilities of two kinds. Birnbaum was apparently unable to express this idea more precisely, e.g., as an equality between two 'evidential meanings'. Nevertheless he goes on to argue that U is *incompatible* with C (conditionality) and thus with L (likelihood). This is the source of the anomaly referred to in the title. It seemed to Birnbaum that there was a broad consensus among statisticians that any adequate conception of statistical evidence should include both U and C. But he had shown that

this is impossible. He concluded the manuscript with the following 'trilemma':

> The only theories which are formally complete, and of adequate scope for treating statistical evidence and its interpretations in scientific research contexts, are Bayesian; but their crucial concept of prior probability remains without adequate interpretation in these contexts. Each of the non-Bayesian alternatives, one identified with the likelihood concept and the other with the error probability concept, seems an essential part of any adequate concept of evidence, but each separately is seriously incomplete and inadequate; however, these cannot be combined because they are incompatible.

The only qualification to this pessimistic conclusion is in the case of 'binary experiments', i.e., experiments in which the parameter space contains only two points. In this case L and U coincide.

The main results of the 1964 manuscript appear in the 1969 paper 'Concepts of Statistical Evidence'. This paper was explicitly intended more for philosophers than statisticians, as is clear from the presentation and the place of publication – in a festschrift for Ernest Nagel. This is Birnbaum's most accessible work. He begins with a positive aim:

> The problem-area of main concern here may be described as that of determining *precise concepts of statistical evidence* (systematically linked with mathematical models of experiments), concepts which are to be *non-Bayesian, non-decision-theoretic*, and significantly *relevant to statistical practice* (Birnbaum [28], p. 113).

The confidence concept is introduced more indirectly and in more general terms than in the earlier manuscript. There is no mention of criterion U. But the equivalence of C with L and the incompatibility of both with the confidence concept are developed in detail. Referring to the existence of these relations, he writes:

> This has surprised and disappointed some, including this writer, who remain without an adequate precise general concept of statistical evidence, and without even consistent criteria for adequacy for such a concept (Birnbaum [28], p. 131).

Although its positive aim was not realized, this paper contains much of great value to anyone interested in statistical inference, and inductive inference generally.

4. BEHAVIORAL VS. EVIDENTIAL INTERPRETATIONS OF THE CONFIDENCE CONCEPT

At the beginning of 'Concepts of Statistical Evidence', Birnbaum distinguished three approaches to statistical inference: (i) Bayesian, (ii) non-Bayesian, decision theoretic, and (iii) non-Bayesian, non-decision theoretic. In this 1969 paper, the primary emphasis is on distinguishing Bayesian approaches from the non-Bayesian, non-decision theoretic approach based on the confidence concept. This is done by emphasizing the incompatibility between the confidence concept and the likelihood principle which is a direct consequence of the Bayesian approach. In Birnbaum's last paper the objective is to distinguish the decision theoretic from non-decision theoretic approaches. The distinction is drawn in terms of two 'interpretations' of the confidence concept which makes its first explicit public appearance.

(Conf) A concept of statistical evidence is not plausible unless it finds "strong evidence for H_2 as against H_1" with small probability (α) when H_1 is true, and with much larger probability ($1-\beta$) when H_2 is true (Birnbaum [41]).

Note that this version takes account of *both* error probabilities.

The confidence concept seems to be central to the Neyman-Pearson account of hypothesis testing and interval estimation. But as these methods were generalized by Neyman and Wald into a full-fledged statistical decision theory, the result of a test was no longer a statement about a hypothesis, but a decision to take some particular course of action. Thus the standard example of acceptance sampling in industrial quality control. Following Neyman's terminology, Birnbaum calls this the 'behavioral' interpretation of the confidence concept. He advocates an 'evidential' interpretation in which the result of a test is not literally an action, but a relative evidential evaluation of two or more hypotheses.

It turns out that a crucial difference between these two interpretations of the confidence concept is that the behavioral interpretation admits a mixture of tests as a legitimate test. Birnbaum asserts that the evidential meaning of a mixed text may be different than that of a single experiment with the same formal characteristics, i.e., the same error probabilities.

This sounds like some version of the principle of conditionality, though this notion is not explicitly employed. Since we know that at least some versions of the principle of conditionality are incompatible with the confidence concept, the question naturally arises whether it is possible to give a precise and logically consistent rendering of the evidential interpretation of the confidence concept. This remains an open question.

The key to Birnbaum's evidential interpretation of the confidence concept is his notion of a *research situation*. All his judgments that some result provides strong or weak evidence, or that some test is to be preferred to another, refer to some hypothetical research situation. The notion of a research situation remains undeveloped, but this clearly is the direction in which he was moving. Probably the best source for his thoughts on what constitutes a scientific research context is the paper 'The Random Phenotype Concept' (Birnbaum [38]). The original manuscript of this paper contained a long section on statistical methodology which, unfortunately, was cut out by the editors of *Genetics*. But what remains gives a fair idea of what Birnbaum regarded as good statistical practice.

5. BIRNBAUM'S PERSPECTIVE FOR RESEARCH IN STATISTICS

Birnbaum distinguished *theoretical statistics* from mathematical statistics, which he regarded as primarily a branch of mathematics like probability theory or measure theory. Theoretical statistics lies between mathematical statistics and substantive science. The job of the theoretical statistician is to develop statistical methods for particular types of scientific problems, and to do this well one must know *both* the relevant mathematics and the relevant science. For Birnbaum this conception of statistics had implications stretching all the way from how statistics should be taught to how one should proceed to resolve the most abstract meta-questions about the nature of statistical evidence. These implications were set out explicitly in a short paper entitled 'A Perspective for Strengthening Scholarship in Statistics' (Birnbaum [34]).

The way to pursue the outstanding issues in the foundations of statistics, Birnbaum maintained, is through the examination of 'case studies' in science. This should be done in an interdisciplinary fashion by statisticians, scientists and philosophers of science. If the case study is historical,

it should involve historians of science as well. Birnbaum was particularly interested in genetics. He studied Mendel's paper in great detail. His own paper on the random phenotype concept (Birnbaum [38]) was explicitly designed as a contemporary case study exhibiting the use of non-Bayesian, non-decision theoretical methods of 'data analysis' in a specific scientific context. His hope was that others with different views would apply their methods to the same problem so that the operative differences might be clearly exhibited and discussed. Of course there is no guarantee that such comparisons will lead to the resolution of the differences, but it is hard to deny that it could be very helpful. It is certainly worth trying.

Indiana University

NOTES

* During the academic year 1971–72 I was Allan Birnbaum's research associate at New York University where he was on the faculty of the Courant Institute of Mathematical Sciences. I shall always be grateful for that experience. We exchanged views almost daily on all manner of issues in philosophy and the foundations of statistics. The following year he left for England where he was a visitor first at Cambridge and later at the University of London. He then accepted a chair at the City University of London. I never saw him again. We exchanged several papers and some letters, but I have little detailed knowledge of what he was doing or thinking while he was in England.

[1] In copies of this letter which Birnbaum himself circulated, the word 'now' is scratched out. He was obviously unhappy with the implication that he had once been an exponent of the likelihood principle.