1. Summary

The paper was presented at a special discussion meeting of the ASA on December 27, 1961 in New York City. It studies the likelihood principle (LP) and how the likelihood function can be used to measure the evidence in the data about an unknown parameter. Essentially, the LP says that if two experiments about the same unknown parameter produce proportional likelihood functions, then the same inference should be made in the two cases.

The LP developed mainly from the ideas of Fisher and Barnard. It caught widespread attention in 1962 when Birnbaum showed that the LP was a consequence of the more accepted principles of sufficiency (that a sufficient statistic summarizes all evidence from an experiment) and conditionality (that experiments not actually performed are irrelevant to inference). Since then the LP has been debated extensively with regard to its place in the foundations of statistical theory. The radical consequences of the LP to statistical analysis have been discussed in many areas. Two of the most important implications are that at the inference stage, stopping rules and sampling designs in survey sampling are irrelevant.

By far, most of the work on the LP's implications and applications has appeared after the 62-paper. During the last 20 years or so, likelihood approaches have been proposed in areas like (1) estimation with nuisance parameters, (2) prediction, (3) survey sampling, (4) missing data problems, and (5) meta-analysis. The monograph by Berger and Wolpert (1984) (hereafter denoted BW) gives an extensive and incisive presentation of the LP, discussing validity, generalizations, and implementations. Otherwise, general discussions of the LP and its consequences can be found in Cox and Hinkley (1974), Basu (1975), Dawid (1977), Dawid (1981), and Barnett (1982).
2. Impact of the Paper

Birnbaum’s main result, that the LP follows from (and implies) sufficiency and conditionality principles that most statisticians accept, must be regarded as one of the deepest theorems of theoretical statistics, yet the proof is unbelievably simple. The result had a decisive influence on how many statisticians came to view the likelihood function as a basic quantity in statistical analysis. Still, even though the impact of this result alone has made a major contribution to the theory of statistics as illustrated in Sec. 5, the paper’s contribution is not limited to this fundamental achievement. It has also affected in a general way how we view the science of statistics. Birnbaum introduced principles or axioms of equivalence within and between experiments, showing various relationships between these principles. This made it possible to discuss the different concepts from alternative viewpoints, thereby discovering weaknesses and strengths of the concepts. Birnbaum’s approach also meant that various statistical “philosophies” could be discussed on a firm theoretical basis. Hence, the paper changed our way of thinking about statistical theories, giving all of us a most important and lasting contribution whether we agree with the LP or not.

3. The Development of Likelihood Ideas

Prior to 1962

At the time when Birnbaum’s paper appeared in 1962, likelihood ideas and methods did not attract much attention in the statistical community. The Neyman–Pearson school and Wald’s decision theory were the dominating approaches, also for statistical problems not of a decision theoretic nature.

The major proponents of likelihood-based inference before Birnbaum’s paper were Fisher and Barnard. Fisher’s theory of estimation (excluding fiducial interval estimation) was essentially a pure likelihood approach, developed in the papers of 1922, 1925, and 1934. Barnard (1947) gave the first version of the LP.

1. Fisher’s Contributions

The term “likelihood” first appeared in Fisher (1921) where the different nature between likelihood and probability was emphasized. The likelihood function as a basis for estimation was introduced by Fisher (1922) when the concepts of information and sufficiency and the method of maximum likelihood were presented. Here, Fisher also used the likelihood to measure the relative support the data give to different values of the parameter. When this paper appeared, the Bayesian theory of Laplace was the main approach to statistical inference. Fisher’s likelihood-based alternative, together with his sharp criticism of the use of prior distributions, especially priors to represent ignorance [see Fisher (1932, p. 258)], led to a lesser interest in Bayesian inference.

After 1922, Fisher’s work on the foundations of statistical inference emphasized likelihood and conditionality, in particular, in Fisher (1925, 1934). The concept of likelihood-based information played a central role in Fisher’s estimation theory and in the development of conditional inference.

Although he came close to asserting a principle of likelihood inference in the theory of estimation [see Fisher (1973, p.73) where he states that in the theory of estimation it has appeared that the whole of the information is comprised in the likelihood function], it seems Fisher never actually stated the likelihood principle in general and may not have been thinking of it as a separate statistical principle [see Fraser (1976)]. Fisher did, however, state a conditionality principle when in 1934, the theory of exact conditional inference based on ancillary statistics was developed for translation families. Fisher’s conditionality principle was motivated by finding the right measure of precision for the maximum likelihood estimate (m.l.e.), and this was attained by conditioning on the maximal ancillary statistic. This conditioning also recovered the information lost by the m.l.e. It should be noted that an ancillary statistic, as used by Fisher, was by definition a part of the minimal sufficient statistic and an index of precision of the m.l.e., not just any statistic whose distribution is independent of the parameter. Since Fisher also supported the principle of sufficiency, his theory of estimation, in effect, agreed with the LP. In general, however, Fisher did not follow the LP. For example, in tests of significance, he advocated using the P-value [see Fisher (1973)], which violates the LP.

Fisher, in his 1956 book, also proposed a likelihood function for prediction [see Fisher (1973, p. 135)]. More than 20 years would pass before the idea of likelihood-based prediction was developed further by Hinkley (1979).

2. Other Contributors

The first formal statement of a version of the likelihood principle was by Barnard (1947, 1949) in the case where the parameter has only two possible values and the LP reduces to stating that two experiments with the same likelihood ratio should lead to the same decision. This was in strong disagreement with the Neyman–Pearson theory that at the time was the dominant approach to hypothesis testing.

Barnard (1949) argued for the LP from an abstract theory of likelihood concepts based on log (odds) instead of probability. From this theory, it was deduced that the relevant measure of evidence of one parameter value \( \theta \) against another \( \theta' \) is the likelihood ratio. It was also shown, under certain assumptions, from the frequentist point of view (with the usual probability
model as a basis) that in choosing between \( \theta \) and \( \theta' \), the decision must be based on the likelihood ratio.

Likelihood concepts were also employed by several other statisticians. Some references are listed in Kendall (1946, pp. 45, 83). For example, Bartlett (1936) used conditional and marginal likelihood for estimating one parameter in the presence of nuisance parameters, and Bartlett (1953) considered approximate confidence intervals based on the derivative of the log likelihood.

A conditionality principle plays a major role in Birnbaum’s paper. A weaker version of this principle appeared in Cox (1958). Cox challenged several of the usual frequentistic approaches to inference, emphasizing the importance of conditioning with several illuminating examples. Cox’s view on conditioning seems to have been essentially the same as Fisher’s.

4. Contents of the Paper

The main aim of the paper is to show and discuss the implication of the fact that the LP is a consequence of the concepts of conditional frames of reference and sufficiency. To this aim, principles of sufficiency, conditionality, and likelihood are defined in terms of the concept of the evidential meaning of an outcome of an experiment.

A second aim of the paper is to describe how and why these principles are appropriate ways to characterize statistical evidence in parametric models for inference purposes. The paper is concerned primarily with approaches to inference that do not depend on a Bayesian model.

After the introduction, the paper is divided into two parts. Part I deals with the mentioned principles of inference and how they relate to each other. Part II deals mainly with how the likelihood function can be interpreted as a measure of evidence about the unknown parameter and considers how commonly used concepts like significance levels and confidence levels can be reinterpreted to be compatible with the LP, leading to intrinsic methods and levels. A discussion section follows the paper. The discussion shows that the participants are very much aware of the vast implications for statistical inference of adopting the LP.

In summarizing the contents of the paper, the sections are named as in the paper.

1. Introduction, Summary, and General Conclusions

An experiment \( E \) is defined as \( E = \{\Omega, S, f(x, \theta)\} \), where \( f \) is a density with respect to a \( \sigma \)-finite measure \( \mu \) and \( \theta \) is the unknown parameter. \( \Omega \) is the parameter space and \( S \) the sample space of outcomes \( x \) of \( E \). The likelihood function determined by an observed outcome is then \( L_x(\theta) = f(x, \theta) \).

Birnbaum restricts attention to problems of informative statistical inference, where one is interested in summarization of evidence or information about \( \theta \) as provided by \( E \) and \( x \) alone [denoted by \( E(\theta, x) \)], and distinguishes two main problems of informative inference: (1) principles that statistical evidence should follow (part I) and (2) interpretation of statistical evidence in accordance with accepted principles (part II). The principles all prescribe conditions under which we should require the same inference for \( (E, x) \) and \( (E', x') \).

The introduction summarizes the principles of sufficiency (S), conditionality (C), and likelihood (L) defined in Secs. 3–5. (S) and (C) are derived from Fisher’s ideas on sufficiency and ancilarity. Birnbaum gives the following formal LP:

\[
(L) \text{ Let } E \text{ and } E' \text{ be two experiments (with common parameter } \theta). \text{ Assume } x, y \text{ are outcomes of } E, E' \text{ respectively with proportional likelihood functions, i.e., } L_{\theta}(x) = c L_{\theta}(y) \text{ for all } \theta \in \Omega, \text{ for some constant } (\text{in } \theta) c > 0. \text{ Then: } E(\theta, x) = E(\theta, y).
\]

Note that the case \( E = E' \) is included. Birnbaum states, without proof, that (L) implies that \( E(\theta, x) \) depends on \( E \) and \( x \) only through \( L_{\theta}(\theta) \). A proof of this result in the discrete case can be found in BW (p. 28). The three principles are described informally in the following way: (S) asserts the irrelevance of observations independent of a sufficient statistic. (C) asserts the irrelevance of experiments not actually performed. (L) asserts the irrelevance of outcomes not actually observed.

This section concludes with stating the main result of the paper [(S) and (C) together are equivalent to (L)] and a discussion of the radical consequences of (L) to the theory and practice of statistical inference. One aspect of the main result is that the likelihood function is given new support, independent of the Bayesian point of view.

Part I

2. Statistical Evidence

This section introduces the concept of the evidential meaning of an observation from an experiment and discusses the term informative inference, where Cox (1958) is a main reference.

Birnbaum states that the central purpose of the paper is to clarify the essential structure and properties of statistical evidence, termed the evidential meaning of \( (E, x) \) and denoted by \( E(\theta, x) \), in various instances. We can say that \( E(\theta, x) \) is the evidence about \( \theta \) supplied by \( x \) and \( E \). Nothing is assumed about what \( E(\theta, x) \) actually is. It can be a report of the experimental results, the inferences made, the methods used, or a collection of different measures of evidence. Birnbaum restricts attention to problems of informative statisti-
3. The Principle of Sufficiency

Here a principle of sufficiency (S) is defined in terms of \( Ev(E, x) \):

\[
(S): \text{Let } t(x) \text{ be a sufficient statistic for } E, \text{ and let } E' \text{ be the experiment of observing } t = t(x). \text{ Then: } Ev(E, x) = Ev(E', t(x)).
\]

The following result is shown by using a result from Bahadur (1954):

**Lemma 1.** Assume \( x, x' \) are two outcomes of \( E \) with proportional likelihood functions. Then: \( (S) \Rightarrow Ev(E, x) = Ev(E, x') \).

We note that this does not mean that (S) implies (L) since \( x, x' \) are from the same experiment.

4. The Principle of Conditionality

This section considers conditional frames of reference and defines a principle of conditionality (C) in terms of \( Ev(E, x) \), first stated in Birnbaum (1961) and related to a discussion by Cox (1958) who provided the crucial idea of a mixture experiment. According to Birnbaum [in a discussion of Barnard et al. (1962)], it was this paper by Cox that made him appreciate the significance of conditionality concepts in statistical inference. The principle is as follows:

\[
(C): \text{Let } E \text{ be a mixture of experiments with components } \{E_k\} \text{ (with common unknown } \theta), \text{ where } E_k \text{ is selected by a known random mechanism. I.e., } E \text{ consists of first selecting a component experiment } E_k \text{ and then observing the outcome } x_k \text{ of } E_k \text{ such that the outcome of } E \text{ can be represented as } (h, x_k). \text{ Then: } Ev(E, (h, x_k)) = Ev(E, x_k).
\]

(C) asserts that the experiments not actually performed are irrelevant. It is stated, without proof, that \( (C) \Rightarrow (S) \). This is not correct. Birnbaum (1972) considers the discrete case and shows that (S) is implied by (C) and the principle of mathematical equivalence (M) that states: If \( x \) and \( x' \) are two outcomes of the same experiment \( E \) with identical likelihood functions, then \( Ev(E, x) = Ev(E, x') \).

The main part of this section serves to illustrate, through various examples, why (C) seems unavoidable in interpreting evidence. It is shown that (C) alone implies quite radical consequences for inference theory, and Birnbaum believes that (C) will be generally accepted. The only serious criticism of (C) in the literature seems to have been Durbin (1970). BW (p. 45) illustrates why Durbin's objection is not convincing. (C) or some version of it seems today to be accepted by most statisticians of various schools of inference.

5. The Likelihood Principle

This section contains the main result of the paper.

**Lemma 2.** (S) and (C) \( \Leftrightarrow (L) \).

The proof is surprisingly simple. Because the proof itself of \( (\Rightarrow) \) has played an important part in the discussion of this result, we shall give the reader a brief outline. Let \( (E_1, x_1) \) and \( (E_2, x_2) \) have proportional likelihood functions and construct then the mixture experiment \( E^* \) that chooses \( E_1 \) with probability \( 1/2 \). Then from (C), it follows that it is enough to show that \( Ev(E^*, (1, x_1)) = Ev(E^*, (2, x_2)) \), which follows from (S) since \( (1, x_1) \) and \( (2, x_2) \) have proportional likelihood functions in \( E^* \).

The implication \( (\Rightarrow) \) is the most important part of the equivalence, because this means that if you do not accept (L), you have to discard either (S) or (C), two widely accepted principles. The most important consequence of (L) seems to be that evidential measures based on a specific experimental frame of reference (like P-values and confidence levels) are somewhat unsatisfactory (in Birnbaum's own words). In other words, (L) eliminates the need to consider the sample space or any part of it once the data are observed. Lemma 2 truly was a "breakthrough" in the foundations of statistical inference and made (L) stand on its own ground, independent of a Bayesian argument. As Savage (1962) noted in his discussion of the paper,

Without any intent to speak with exaggeration or rhetorically, it seems to me that this is really a historic occasion. This paper is a landmark in statistics because it seems improbable to me that many people will be able to read this paper or to have heard it tonight without coming away with considerable respect for the likelihood principle.

Part II

6. Evidential Interpretations of Likelihood Functions

This is a short section describing the purposes of Sec. 7–9 that are mainly concerned with the question: What are the qualitative and quantitative properties of statistical evidence represented by \( L_\theta(\theta) \)?

7. Binary Experiments

This section covers the case \#(\Omega) = 2 and is closely related to parts of Birnbaum (1961). Let \( \Omega = (\theta_1, \theta_2) \). In this case, (L) means that all information
lies in the likelihood ratio, \( \lambda(x) = f(x, \theta_2)/f(x, \theta_1) \). The question is now what evidential meaning [in accordance with (L)] we can attach to the number \( \lambda(x) \). To answer this, Birnbaum first considers a binary experiment in which the sample space has only two points, denoted (+) and (−), and such that \( P(+|\theta_1) = P(−|\theta_2) = \alpha \) for all \( \alpha \leq 1/2 \). Such an experiment is called a symmetric simple binary experiment and is characterized by the "error" probability \( \alpha \). For such an experiment, \( \lambda(+) = (1 - \alpha)/\alpha \) and \( \alpha = 1/(1 + \lambda(+)) \). The important point now is that according to (L), two experiments with the same value of \( \lambda \) have the same evidential meaning about the value of \( \theta \). Therefore, the evidential meaning of \( \lambda(x) \geq 1 \) from any binary experiment \( E \) is the same as the evidential meaning of the (+) outcome from a symmetric simple binary experiment with \( \alpha(x) = 1/(1 + \lambda(x)) \). \( \alpha(x) \) is called the intrinsic significance level and is a measure of evidence that satisfies (L), while usual observed significance levels (P-values) violate (L). \( 1 - \alpha(x) \) is similarly called the intrinsic power at \( x \).

8. Finite Parameter Spaces

Section 8.1 illustrates that some likelihood functions on a given \( \Omega \) can be ordered in a natural way by constructing equivalent experiments with sample spaces consisting only of two points.

Let \( k = \#(\Omega) \). In Sec. 8.2, intrinsic confidence methods and intrinsic confidence levels for an outcome \( x \) are defined. This is done in a similar fashion as in Sec. 7 by constructing an experiment \( E' \) with \( \#(S) = k \) based on \( L_a(\theta) \) such that the likelihood function for one outcome in \( E' \) is equal to \( L_a(\theta) \). Then intrinsic confidence methods and levels are defined as regular confidence methods and levels in \( E' \).

Sections 7 and 8 show that for finite parameter spaces, significance levels, confidence sets, and confidence levels can be based on the observed \( L_a(\theta) \) [hence, satisfying (L)], defined as regular such methods and constructed for a constructed experiment with a likelihood function identical to \( L_a(\theta) \). Therefore, in the case of finite parameter spaces, a clear and logical evidential interpretation of the likelihood function can be given through intrinsic methods and concepts.

9. More General Parameter Spaces

This section deals mainly with the case where \( \Omega \) is the real line. Given \( E, x \), and \( L_a(\theta) \), a location experiment \( E' \) consisting of a single observation of \( Y \) with density \( g(y, \theta) \propto L_a(\theta - y) \) is then constructed. Then \( (E, x) \) has the same likelihood function as \( (E', 0) \), and (L) implies that the same inference should be used in \( (E, x) \) as in \( (E', 0) \). For example, if a regular \( (1 - \alpha) \) confidence interval in \( E' \) is used, then this interval estimate (for \( y = 0 \)) should be the one used also for \( (E, x) \) and is called a \( (1 - \alpha) \) intrinsic confidence interval for \( (E, x) \). There is, however, one major problem with this approach: A nonlinear transformation of \( \theta \) will lead to a different \( g(y, \theta) \) and hence different intrinsic statements. This problem does not arise in Sec. 7 and 8, where \( \Omega \) is finite.

Birnbaum considers the case where \( L_a(\theta) \) has the form of a normal density and defines as an index of the precision of the maximum likelihood estimate, \( \hat{\theta}(x) \), the standard deviation in \( g(y, \theta) \), calling it the intrinsic standard error of \( \hat{\theta}(x) \). Of course, according to (L), the usual standard error is not a proper measure of precision.

As a general comment, Birnbaum emphasizes that intrinsic methods and concepts can, in light of (L), be nothing more than methods of expressing evidential meaning already implicit in \( L_a(\theta) \) itself. In the rejoinder in the discussion, Birnbaum does not recommend intrinsic methods as statistical methods in practice. The value of these methods is conceptual, and the main use of intrinsic concepts is to show that likelihood functions as such are evidentially meaningful.

Sequential experiments are also considered, and it is noted that (L) implies that the stopping rule is irrelevant (the stopping rule principle).

10. Bayesian Methods: An Interpretation of the Principle of Insufficient Reason

Birnbaum views the Bayes approach as not directed to informative inference, but rather as a way to determine an appropriate final synthesis of available information based on prior available information and data. It is observed that in determining the posterior distribution, the contribution of the data and \( E \) is \( L_a(\theta) \) only, so the Bayes approach implies (L). In this section, the case of uniform priors to represent the absence of prior information is discussed.

11. An Interpretation of Fisher's Fiducial Argument

The aim of Fisher's fiducial approach is the same as the Bayes approach: Statements of informative inference should be in the form of probabilities. However, no ignorance priors are involved. However, no ignorance priors are involved. In fact, as mentioned earlier, Fisher argued strongly against the use of noninformative priors. Birnbaum suggests that the frames of reference in which fiducial probabilities are considered may coincide, in general, with the constructed experiments in which intrinsic methods are defined and hence that fiducial confidence methods may coincide with intrinsic confidence methods. He formulates a fiducial argument compatible with (L) and shows how, in the case of finite \( \Omega \), this modified fiducial approach corresponds to the intrinsic approach in Sec. 8. However, Birnbaum's fiducial argument seems hard to generalize to the usual type of parameter spaces.

12. Bayesian Methods in General

In this section, Birnbaum considers what may separate Bayesian methods (with proper informative priors) from methods based only on \( L_a(\theta) \). It is
observed that in binary experiments, the class of "likelihood methods" is identical to the class of Bayesian methods for the problem of deciding between the two parameter values $\theta_1, \theta_2$.

13. Design of Experiments for Informative Inference

One may think that (L) has nothing to say in design problems since no data and hence no likelihood function are available. However, as Birnbaum mentions in this section, according to (L), the various experimental designs are to be evaluated and compared only in terms of inference methods based solely on the likelihood functions the designs determine, along with the costs. Illustrations are given using intrinsic methods.

5. Developments After Birnbaum's Paper

1. Discussion of Lemma 2, (S) and (C) $\Leftrightarrow$ (L)

Lemma 2 has been widely discussed in the literature and, as stated, is relevant only in the discrete case. Also, the proof of (L) $\Rightarrow$ (S) was not stated correctly. A correct proof in the discrete case is given by Birnbaum (1972). Birnbaum (1961) had shown that (S) and (C) $\Rightarrow$ (L) for binary experiments (i.e., $\Omega$ consists of two points only).

In nondiscrete cases, there are some problems. Joshi (1976) raised objections to the definition of (S) in the continuous case and showed that a trivial application of (S) would suggest that in a continuous model, $Ev(E, x)$ is identical to the same for all $x$. BW (Sec. 3.4.3) considers the nondiscrete case generally and suggests modifications of (C), (S) and (L) that lead to the same implication of (C) and (S). However, as Basu (1975) notes, the sample space in any realizable experiment must be finite due to our inability to measure with infinite precision, and continuous models are to be considered as mere approximations. One may therefore consider arguments for (L) in the discrete case as all that is needed in order to argue for (L) as a principle for inference in all experiments [also, Barnard et al. (1962) and Birnbaum (1972) discuss this point].

Note that Lemma 2 applies only to experiments that can be represented completely and realistically by the form $E = \{\Omega, S, f(x, \theta)\}$. Savage, Barnard, Bross, Box, Levene, and Kempthorne all commented on this aspect in the discussion of the paper. Later, objections to this representation of an experiment have been raised in the theories of pivotal and structural inference; see, e.g., Barnard (1980) and Fraser (1972). Birnbaum in his rejoinder discusses the possibility of a likelihood approach to robustness problems by enlarging the parametric model to a class of models labeled by nuisance parameters.

Basu (1975) considered the discrete case and defined the following weaker versions of (S) and (C):

**Weak sufficiency principle** [named by Dawid (1977)]: (S'): Let $t(x) = t(x')$. Then $Ev(E_x, x) = Ev(E_{x'}, x')$.

**Weak conditionality principle** [named by Basu]: (C'): Let $E$ be a mixture of $E_1$ and $E_2$ with known mixture probabilities $\pi$ and $1 - \pi$. Then $Ev(E, (x, x_2)) = Ev(E_{x_2}, x_2)$.

Basu recognized that the proof of Lemma 2 requires only (S') and (C') and that Birnbaum in fact showed (S') and (C') $\Leftrightarrow$ (L). Actually, Birnbaum's proof shows that this result is true for (C') with $\pi = 1/2$.

Statisticians using sampling-theory-based inference do not act in accordance with (L) and must reject or at least modify (S), (S') or (C), (C'). Durbin (1970) and Kalbfleisch (1975) attempt such modification. Durbin suggests that in (C), the ancillary statistic $h$ must depend on the minimal sufficient statistic. It is shown that the proof of Lemma 2 fails when the domain of (C) is restricted in this way. Arguments against Durbin's suggestion have been made by Savage (1970), Birnbaum (1970), and BW. As mentioned earlier, an example given by BW (p. 45) illustrates that Durbin's restriction seems unreasonable. Kalbfleisch (1975) distinguishes between experimental and mathematical ancillaries (an experimental ancillary is determined by the experimental design, and a mathematical ancillary is determined by the model of the problem) and suggests that (S) [or (S')] should apply only to experiments with no experimental ancillaries. Then (L) does not follow from (C) and (S).

Kalbfleisch's suggestion was criticized in the discussion of his paper, especially by Birnbaum and MacLaren. The main problems with such a restriction of sufficiency are (as mentioned by BW, p. 46) the following (1) It seems artificial to restrict principles of inference to certain types of experiments. (2) It is difficult to distinguish between mixture and nonmixture experiments. (3) Mixture experiments can often be shown to be equivalent to nonmixture experiments [Birnbaum and MacLaren illustrate (2) and (3)]. (4) In almost any situation, behavior in violation of sufficiency can be shown to be inferior. Joshi (1990) claims there is an error in the proof of Lemma 2 in the discrete case (as presented in BW, p. 27), but as made clear by Berger (1990) in his response, the proof is in fact correct. Joshi simply argues for the same type of restriction of sufficiency as Kalbfleisch (1975).

By restricting consideration to the discrete case, various alternative principles to (C') and (S') also imply (L). Birnbaum (1972) showed that (M) and (C') $\Leftrightarrow$ (L). (M) was scrutinized by Godambe (1979) who disagreed with Birnbaum's interpretation of it. Pratt (1962) advanced an alternative justification of (L) based on a censoring principle (Ce) and (S). (Ce) was formalized by Birnbaum (1964) and is given by

(Ce) For a given experiment $E$ with sample space $S$, let $E^*$ be a censored version of $E$ with sample space $S^*$ such that certain points in $S$ cannot be observed. If $x$ is an outcome in both $S$ and $S^*$, then $Ev(E_x, x) = Ev(E^*_x, x)$. 

*Introduction to Birnbaum (1962)*
Birnbaum (1964) proved that (Ce) and (5) imply (L) [see also Dawid (1977)] and in his 1972 paper finds (Ce) simpler and at least as plausible as (C). Dawid (1977) and Berger (1984) show that also other principles lead to (L).

2. Other Developments

Another major paper on likelihood inference, Barnard et al. (1962), appeared right after Birnbaum's paper and was read before the Royal Statistical Society in March 1962. They stated the same likelihood principle as Birnbaum in terms of making an inference about \( \theta \), and tried to argue that (L) follows from (S'). However, the argument was fallacious, as shown by Armitage and Birnbaum in the discussion of the paper. Likelihood methods consisting essentially of plotting the whole likelihood function were proposed and applied to autoregressive series, Markov chains, and moving average processes. As Birnbaum did, they showed that the stopping rule principle (SRP) is a consequence of (L). Pratt (1965) also discusses the SRP. A general study of the SRP is given by BW (Sec. 4.2).

Various examples have been constructed with the aim of showing that the likelihood principle leads to unreasonable inferences. These include "the stopping rule paradox," first discussed it seems by Armitage (1961), and the examples of Stein (1962), Stone (1976), and Joshi (1989). BW and Hill (1984) discuss the first three of these examples (and also a version of Joshi's example), Basu (1975) considers Stein's example and "the stopping rule paradox," and Good (1990) examines the example by Joshi. They all argue essentially that none of these examples speak against (L) itself, but rather against certain implementations of (L).

During the last three decades, the implementation of (L) by considering methods based on the observed likelihood function only (non-Bayesian likelihood methods) has been considered by many statisticians. We have already mentioned, in addition to Birnbaum, Barnard et al. (1962). Most of the likelihood methods that have been proposed depend on the interpretation that \( L_\alpha(\theta_1)/L_\alpha(\theta_2) \) measures the relative support of the data for \( \theta_1 \) and \( \theta_2 \). Development of this idea can be found in Hacking (1965) and Edwards (1972). In the case of nuisance parameters, likelihood approaches have been suggested by, among others, Kalbfleisch and Sprott (1970), Sprott (1975), Cox (1975), Dawid (1975), Barndorff-Nielsen (1978), Barndorff-Nielsen (1986), Cox and Reid (1987), Fraser and Reid (1988), and McCullagh and Tibshirani (1990). Barnard (1967) discusses the use of the likelihood function in inference with applications to particle physics and genetics. Rubin (1976) considers likelihood-based inference for missing data problems [see also Little and Rubin (1987)]. Goodman (1989) suggests a likelihood approach to meta-analysis (the science of combining evidence from different studies) based on log-likelihood ratios. Other references can be found in BW (Sec. 5.2).

In maximum likelihood estimation, the expected Fisher information as a precision index of the estimate is not appropriate according to (L). The suggestion by Fisher (1925, 1934) of using instead the observed Fisher information [named by Edwards (1972)] does satisfy (L) and is also supported from a frequentist point of view as shown by Hinkley (1978) and Efron and Hinkley (1978).

Several writers have discussed the fact that (L) leads to a rejection of significance tests as valid measures of inferential evidence. We refer to BW (Sec. 4.4) for references.

One of the areas where the LP has had a major impact is in survey sampling. Two of the most important implications of (L) to survey sampling are: (1) : (L) \( \Rightarrow \) sampling design is irrelevant at the inference stage, and (2) : (L) \( \Rightarrow \) modeling of the population. In the next two paragraphs, we outline the development of these two implications.

(1) It was first shown by Godambe (1966) and Basu (1969) that, with usual noninformative sampling design, the likelihood function is flat for all possible values of the parameter (the population vector). Hence, (L) implies that the inference should not depend on the sampling design. However, Godambe (1966, 1982) claims that (L) may not be appropriate here since there is a relationship between the parameter and data (which is a part of the parameter.) It should be noted that Lemma 2 in Birnbaum's paper is valid also in this case, of course. Basu does not find Godambe's argument convincing and concludes that the sampling design is irrelevant at the inference stage. This was in dramatic disagreement with the classical approach.

(2) The fact that the likelihood function is flat have by some been viewed as a failure of (L) [see, e.g., Rao (1971)], but can also be seen as clarifying some of the limitations of the conventional model as noted by Royall (1976). Royall (1976) seems also to have been the first to recognize that (L) makes it necessary in a sense to model the population (see also BW, p. 114). From the likelihood principle point of view, the data do not, in fact, contain any information about the unobserved part of the population. To make inference, it is therefore necessary to relate the data to the unobserved values somehow, and a natural way of doing this is to formulate a model. Also, as noted by Royall (1971), modeling the population is as objective as any modeling usually done in statistics. General discussions and applications of population modeling can be found in Smith (1976) and Thomsen and Tesfa (1988).

Prediction is another area where a likelihood approach has been attempted. Kalbfleisch (1971) and Edwards (1974) considered Fisher's suggestion of a likelihood function for prediction and Hinkley (1979) coined the term "predictive likelihood" suggesting several such likelihoods. Since then several papers have appeared on the subject. A list of references can be found in Björnstad (1990).

The problem of nonresponse in survey sampling represents a prediction case where predictive likelihood may give valuable contributions. Little (1982) considers some likelihood aspects of the nonresponse problem.

Finally, it should be mentioned that Birnbaum (1968, 1977) came to view
(L) rather critically, because of its conflict with the so-called confidence principle. (For a discussion of this principle see BW, Sec. 4.1.5).

6. Biography

Allan D. Birnbaum was born on May 27, 1923, in San Francisco of Russian Jewish parents. He died in London in July 1976. He studied as an undergraduate at the University of California in Berkeley and Los Angeles, completing a premedical program in 1942 and receiving a bachelor's degree in mathematics in 1945. For the next two years, he took graduate courses at UCLA in science, mathematics, and philosophy. In 1947, Birnbaum went to Columbia where he obtained his Ph.D. in mathematical statistics in 1954 under the guidance of Erich L. Lehmann.

By then, he had been a faculty member at Columbia for three years. He stayed at Columbia until 1959 while also visiting Imperial College, London, and Stanford during this time. In 1959, he moved to the Courant Institute of Mathematical Sciences at New York University, becoming a full professor of statistics in 1963. He remained at the institute until 1972, when he left for an extended visit to Britain. In 1975, he accepted the chair of statistics at the City University of London where he remained until his death.

Birnbaum had several other professional interests, including medicine and philosophy. Four memorial articles about Birnbaum have been published; see Norton (1977), Godambe (1977), Lindley (1978), and Barnard and Godambe (1982).

Acknowledgments

Berger and Wolpert (1984) have been an invaluable help in writing this introduction. I thank the editors for their help in locating some of the memorial articles and some of the recent literature.

References


