More on Concepts of Statistical Evidence

ALLAN BIRNBAUM*

A self-contained account is given of the implications among the sufficiency, conditionality and likelihood axioms of statistical evidence for the discrete case. These include a previously unpublished derivation of sufficiency from conditionality. The nondiscrete case is discussed with reference to the same relation, and to the significance of nonunique determination of density and likelihood functions. The writer's current views on this problem area are indicated briefly.

1. INTRODUCTION AND SUMMARY

Various implication relations among the sufficiency, likelihood, and conditionality axioms are derived in a brief self-contained account restricted to the case of discrete distributions. This constitutes an improvement upon corresponding parts of [1]. In particular the assertion without proof (p. 279) that conditionality implies sufficiency, questioned by Joshi [9], is proved.

Corresponding implications in the nondiscrete case are also discussed, partly with reference to another question raised by Joshi [9] concerning the significance of nonuniqueness of determinations of density and likelihood functions.

No attempt is made here to review the background and significance of these concepts of statistical evidence and their status in relation to statistical practice and theories of inference. My own current general views are no longer represented well by [1], but rather by [3–8], and by brief final comments in this article.

2. AXIOMS AND IMPLICATIONS IN THE DISCRETE CASE

Let E denote any specified mathematical model of an experiment: $E=(\Omega, S, f)$, where $S=\{x\}$ is the discrete sample space, $\Omega=\{\theta\}$ is the parameter space, and $f=f(x, \theta)=\text{Prob}(X=x|\theta)$ denotes the elementary probability function (pdf), for each x, θ. For each x, (E, x) is a model of (an instance of) statistical evidence. A judgment that two such models with common parameter space represent equivalent statistical evidence is represented by writing $Ev(E, x)=Ev(E^*, x^*)$. Only models with the same parameter space will be compared for possible equivalence.

Some readers of [1] and subsequent discussions have found it disconcerting that the range of the function Ev is unspecified and seems rather abstract. It may help to note that Ev is used specifically only to establish relations $Ev(E, x)=Ev(E', x')$, which could alternatively always be written as $(E, x)\sim(E', x')$ where $\sim$ is an equivalence relation to be interpreted as "is evidentially equivalent to." Of course we resort to this indirect approach, considering when it seems right to say that two sets of data have the same significance, only because of difficulties encountered in more direct attempts to characterize in precise general terms "what the data say."

Recall the following *definitions*:

1. $h=h(x)$ is called an *ancillary* statistic if f admits the factored form
$$f(x, \theta) = g(h)f(x \mid h, \theta)$$
where $g=g(h)=\text{Prob}(h(X)=h)$ is independent of θ. Here $f_h=f(x|h, \theta)$ is the indicated conditional pdf. (E_h, x) denotes a model of evidence determined by an outcome x of the experiment E_h: (Ω, S_h, f_h) where $S_h=\{x:h(x)=h\}$. E may be called a mixture experiment, with components E_h having respective probabilities $g(h)$.
2. $t=t(x)$ is a *sufficient* statistic if f admits the factored form
$$f(x, \theta) = g(t, \theta)f(x \mid t)$$
where $g(t, \theta)=\text{Prob}(t(X)=t|\theta)$ and the conditional pdf $f_t=f(x|t)$ is independent of θ. (E', t) denotes a model of evidence determined by outcome t of experiment E': (Ω, S', g) where $S'=\{t:t=t(x), x\in S\}$.

The *axioms of statistical evidence* to be considered are:

Conditionality (C): If $h(x)$ is an ancillary statistic, then $Ev(E, x)=Ev(E_h, x)$, where $h=h(x)$.
Likelihood (L): If, for some $c>0$, $f(x, \theta)=cf^*(x^*, \theta)$ for all $\theta\in\Omega$, then $Ev(E, x)=Ev(E^*, x^*)$.
Sufficiency (S): If $t(x)$ is a sufficient statistic, then $Ev(E, x)=Ev(E', t)$, where $t=t(x)$.
Weak sufficiency (S′): If, for some $c>0$, $f(x, \theta)=cf(x^*, \theta)$ for all $\theta\in\Omega$, then $Ev(E, x)=Ev(E, x^*)$.
Mathematical equivalence (M): If $f(x, \theta)=f(x', \theta)$ for all $\theta\in\Omega$, then $Ev(E, x)=Ev(E, x')$.

Axiom (S′) will be discussed in the Section 3 and is included here for convenience.

Axiom (M) formalizes the simplest case of the "natural" concept (or "obvious" judgment or assertion) that two models of statistical evidence are to be considered equivalent if they differ only in the manner of labelling sample points. A simple example is that of two independent identically distributed Bernoulli trails: The events (0, 1)

* Allan Birnbaum is professor of mathematical statistics, Courant Institute of Mathematical Sciences, New York University, New York, N.Y. 10012, and currently Visiting Fellow, Peterhouse, Cambridge University, Cambridge, England. The author is grateful to Dr. V.M. Joshi for the opportunity to see his papers. The late Professor L.J. Savage very kindly provided several helpful comments on both the form and content of an earlier version of this article including the substance of the second paragraph of Section 2. This research was supported in part by the National Science Foundation and in part by the U.S. Public Health Service.

© Journal of the American Statistical Association
December 1972, Volume 67, Number 340
Theory and Methods Section

and (1, 0) have the same probabilities $\theta(1-\theta)$, for each θ.

In the case where $\theta=.1$ or $.5$ only, we may represent E explicitly in matrix form thus:

$$E = (f(x,\theta)) = (f(j,i)) = \begin{pmatrix} .01 & .09 & .09 & .81 \\ .25 & .25 & .25 & .25 \end{pmatrix}.$$

(M) implies, and is typified by, the judgment that $Ev(E, 2)=Ev(E, 3)$. Such a concept seems implicit in most discussions of the other concepts of evidence considered here. This concept was expressed in [1, p. 278] as part of the basis for general discussion, but was not formalized. The derivations there might be said to depend tacitly on (M), in the sense that they proceed from axioms each logically stronger than (M).

However the derivation of (S) from (C), not included in [1], requires application of such a concept, of which the simple case formalized as (M) suffices for present purposes.

The following theorems will be proved:

Theorem 1: (C) and (M) jointly imply (L).

Theorem 2:

$$\begin{tikzcd}(L) \arrow[r] \arrow[dr] & (S) \arrow[r] & (S') \arrow[r] & (M) \\ & (C) & & \end{tikzcd}$$

Corollary: (C) and (M) jointly imply (S).

Theorem 3: Each of the three implications

$$(L) \leftarrow (S) \leftarrow (S') \leftarrow (M)$$

is false.

The corollary follows immediately from Theorems 1 and 2. The following proof of Theorem 1 is adapted from [2]; the derivation is similar to that of Lemma 2 in [1]. The method indicated in [1, p. 279] is inadequate since it applies just to the restricted (though important) class of examples in which an ancillary and a sufficient statistic are independent. Its inadequacy was shown independently by Joshi [9].

Proof of Theorem 1: Let E and E' denote any two experiments having the common parameter space $\Omega=\{\theta\}$, and represented by pdf's $f(x,\theta)$, $g(y,\theta)$ on their respective sample spaces $S=\{x\}$, $S'=\{y\}$. Let x', y' be any two outcomes of E, E' respectively which determine the same likelihood function; that is, $f(x',\theta)=cg(y',\theta)$ for all θ, where c is some positive constant. Consider the (hypothetical) mixture experiment E^* whose components are just E and E', taken with respective probabilities $k=1/(1+c)$ and $1-k=c/(1+c)$. Denoting by z the generic outcome of E^*, and by $h(z,\theta)$ its pdf, we find

$$h(x',\theta) = kf(x',\theta) = \left(\frac{c}{1+c}\right)g(y',\theta)$$
$$= (1-k)g(y',\theta) = h(y',\theta) \quad \text{for each } \theta.$$

Hence by (M) we have

$$Ev(E^*, x') = Ev(E^*, y'). \tag{2.1}$$

By (C) we have

$$\begin{aligned} Ev(E^*, x') &= Ev(E, x'), \quad \text{and} \\ Ev(E^*, y') &= Ev(E', y'). \end{aligned} \tag{2.2}$$

By (2.1) and (2.2) we have

$$Ev(E, x') = Ev(E', y'),$$

completing the proof.

Proof of Theorem 2: *(L)→(S)*: Suppose t is sufficient in E, and determines the experiment $E'=(\Omega, S', f')$ by the transformation $t=t(x)$, $S'=\{t\}=t(S)$,

$$\text{(A)} \qquad f'(t,\theta) = \textstyle\sum_{t(x)=t} f(x,\theta).$$

As is well known (see e.g., [10, p. 191 ff.]), a statistic $t(x)$ is sufficient in $E=(\Omega, S, f)$ only if $t(x)=t(x')$ implies that for some $c>0$, $f(x,\theta)=cf(x',\theta)$ for all θ. Thus (A) has the form $f'(t,\theta)=cf(x,\theta)$, where $t=t(x)$, for some $c>0$. Thus if $t(x)=t$, the same likelihood function is determined by (E, x) and by (E', t); that is, the latter two models satisfy the hypothesis of (L). Assuming (L), we have $Ev(E, x)=Ev(E', t)$, where $t=t(x)$, which is the conclusion of (S).

(S)→(S'): Suppose that, for some $c>0$, $f(x',\theta)=cf(x'',\theta)$ for all θ. Then (see, e.g., [10, p. 191 ff.]) the statistic t which transforms each point of S into itself, except for x'', which it carries into x', is sufficient. Assuming (*S*), we obtain $Ev(E, x')=Ev(E', t')$ and $Ev(E, x'')=Ev(E', t')$, where $t'=t(x')=t(x'')$. Hence $Ev(E, x')=Ev(E, x'')$, the conclusion of (S').

(S')→(M): Upon setting $c=1$ in (S'), we obtain (M).

(L)→(C): If $h(x)$ is ancillary in E, then $f(x,\theta) = g(h)f(x\,|\,h,\theta)$, where g is a pdf independent of θ. Without essential loss of generality we may assume $g(h)>0$, for each h. Thus the preceding equation has the form of the hypothesis of (L), with respect to (E, x) and (E_h, x). Assuming (L), we obtain $Ev(E, x)=Ev(E_h, x)$ where $h=h(x)$, the conclusion of (C). This completes the proof of Theorem 2.

Proof of Theorem 3: To show that (S) does not imply (L), note that

$$Ev\left(\begin{bmatrix} .1 & .9 \\ .2 & .8 \end{bmatrix}, 1\right) = Ev\left(\begin{bmatrix} .1 & .3 & .6 \\ .2 & .3 & .5 \end{bmatrix}, 1\right)$$

is implied by (L), but not by (S), since no *sufficient* statistic transforms one of these models into the other.

To show that (S') does not imply (S), note that

$$Ev\left(\begin{bmatrix} .01 & .18 & .81 \\ .25 & .50 & .25 \end{bmatrix}, 2\right)$$
$$= Ev\left(\begin{bmatrix} .01 & .09 & .09 & .81 \\ .25 & .25 & .25 & .25 \end{bmatrix}, 2\right)$$

under (S) but not under (S'). (The latter applies only to cases with common E.)

To show that (M) does not imply (S'), note that

$$Ev\left(\begin{bmatrix} .1 & .2 & .7 \\ .2 & .4 & .4 \end{bmatrix}, 1\right) = Ev\left(\begin{bmatrix} .1 & .2 & .7 \\ .2 & .4 & .4 \end{bmatrix}, 2\right)$$

under (S') but not under (M). This completes the proof of Theorem 3.

The role of (M) in the proof of Theorem 1 makes it plausible if not obvious that (C) does not imply (M).

3. THE NONDISCRETE CASE

3.1. In discrete models, if underlying measures more general than the usual counting measure are allowed, then adoption of an axiom of the form of (M), formulated in terms of general densities, is tantamount to adoption of (S′), the principal part of the sufficiency concept. For example, equal densities $f(x, \theta) = f(x', \theta)$, $\theta \in \Omega$, multiplied by positive unequal underlying point measures $\mu(x)$, $\mu(x')$, give *proportional* probabilities $f(x, \theta)\mu(x)$, $f(x', \theta)\mu(x')$, $\theta \in \Omega$, as specified in (S′), which was formulated with reference just to counting measure.

The situation is similar in nondiscrete cases where probabilities are specified by density functions relative to some underlying measure. It appears difficult in this case to formulate and interpret any non-trivial general concept of mathematical equivalence comparable with but weaker than (S′). In this case (S′) and (C) imply (L), as can be proved by the same method of derivation used for Theorem 1, and for Lemma 2 in [1].

3.2. Joshi [9] has questioned both the validity of the derivation of Lemma 2 in [1], and the consistency of applications of the likelihood axiom, on the basis of the well-known nonuniqueness of density functions in the nondiscrete case. Although measure-theoretic considerations were not treated explicitly in [1], they were observed validly and consistently, as the following comments will make clear.

Throughout [1] (beginning on p. 274) it was assumed that a model E of an experiment "is represented by a specified elementary probability function $f(x, \theta)$" with respect to a given underlying measure. In nondiscrete cases, the experimenter can in principle choose (before taking observations) a specific probability density function (among alternative equivalent forms). This is analogous to the choice of any statistic before taking observations; and for our purposes it *is* in fact a choice of a statistic, since the likelihood function is a statistic. The axioms of statistical evidence, and the derivations relating them, thus concern any fixed specified form of a statistic (sufficient; ancillary; likelihood function; or (E, x) itself), as that form may in principle be selected before observation. Thus arbitrariness and inconsistencies among alternative possible selections are limited, as usual in probability and mathematical statistics, to sample points having total probability zero, for each parameter point. Moreover on the basis of any given fixed choice, the derivations and possible applications require *no* qualification, even with respect to any single sample point.

We may illustrate by the example of a sample of one observation normally distributed with unit variance and unknown mean θ. The usual density function (with respect to Lebesgue measure) is

$$f(y, \theta) = \phi(y - \theta) \equiv (1/\sqrt{2\pi}) \exp - \tfrac{1}{2}(y - \theta)^2.$$

This may be replaced, for each respective θ, by

$$f_1(y, \theta) = \begin{cases} \phi(-\theta) & \text{if } y \text{ is rational,} \\ \phi(y - \theta) & \text{otherwise,} \end{cases}$$

without affecting any purpose of probability or mathematical statistics which depends exclusively on probabilities. This change is analogous to replacing the usual (minimal) sufficient statistic y by the statistic

$$t(y) = \begin{cases} 0 & \text{if } y \text{ is rational} \\ y & \text{otherwise.} \end{cases}$$

As is well known, the latter has the same distributions as y, and may thus replace y for all purposes where only probabilities are relevant. In fact $t(y)$ is a one-to-one function of the likelihood function $L_1(\theta, y) = f_1(y, \theta)$, while y is a one-to-one function of the more familiar likelihood function $L(\theta, y) = f(y, \theta)$. (Both y and $t(y)$ are maximum likelihood estimators, with respect to the respective, mathematically equivalent, specifications of the same model.)

4. EXTRA-MATHEMATICAL CONSIDERATIONS

The preceding considerations show that the applied probabilist and the interpreter of research data often may, and sometimes must, choose one among certain equivalent (a.e.μ.) forms of densities and/or statistics. They also show that any possible basis for preference among such alternatives, for example a preference for y rather than $t(y)$ as an estimator, lies outside of mathematical probability theory and mathematical statistics as such. (For a discussion of somewhat comparable questions in other areas of applied mathematics, see [13].)

A preference such as that for y rather than $t(y)$ is sometimes explained as based on a kind of concrete realism. Since actual observations are usually discrete due to rounding off, any density function must be interpreted only as a part of a formula giving approximations to the possible discrete distributions of actual observations. Thus if all outcomes are rounded off to rational numbers y, then (referring to densities defined above) $f(y, \theta)\Delta y$ is a better approximation formula for the probabilities of interest than $f_1(y, \theta)\Delta y$ since at rational y the latter is zero. More generally, such preferences may be expressed by choosing, among possible forms equivalent to $f(y, \theta)$ (a.e.μ.), forms continuous in y and θ jointly.

I believe it is possible to support such preferences also by somewhat distinct and deeper considerations, based on the structure of a scientific discipline, including its empirical and theoretical aspects. However these lie beyond the scope of this article.

Although the preceding sections seem clearly to lie within the scope of mathematical statistics, of course the concepts of statistical evidence themselves and the questions considered in the preceding paragraphs of this section are not subsumed in any mathematical discipline as such but may be better described as parts of *theoretical statistics*, regarded as a discipline concerned with the concepts and problems which link mathematical statistics with applications and interpretations in scientific research and other contexts. (Theoretical statistics in such a sense has been discussed by Tukey, e.g., [14], and in turn by myself, in a partly distinct sense, as invited discussant at the presentation of Tukey's paper.)

My current views concerning concepts of statistical evidence differ appreciably from those in [1] and have been presented in [3–8]. Even from the general standpoint represented by [1], the conditionality concept seems no longer crucial, in the sense that the censoring concept due to Pratt [11, 12, 3] seems at least equally plausible as well as simpler, and has essentially the same implications for statistical theory and practice. (Various implications among the censoring concept and the axioms discussed in the present article were derived in [2].)

More generally, I find that my own interpretations of research data in scientific contexts (such as Mendelian genetics; cf., e.g., [8]) are appropriately developed and formulated with no very systematic reliance on any precise general concept of statistical evidence, but with a limited and informal role for the "confidence concept of statistical evidence." The latter term seems appropriate and convenient to designate the widely current concept by which suitably selected confidence regions and statistical tests are interpreted in research contexts as indicators of statistical evidence. (As Neyman consistently points out, his theories of estimation and testing include no such concept.) This problem area seems to me to be in a state not only unsatisfactory but indeed anomalous, in respects not adequately met by various proposed systematic approaches.

Possible advances may be guided by development of more intensive and critical case studies in research disciplines (cf. [7]) such as Mendelian genetics. Relevant clarifications of the nature and roles of statistical evidence in scientific research may well be achieved by bringing to bear in systematic concert the scholarly methods of statisticians, philosophers and historians of science, and substantive scientists such as geneticists.

[*Received February 1971. Revised March 1972.*]

REFERENCES

[1] Birnbaum, Allan, "On the Foundations of Statistical Inference," with discussion, *Journal of the American Statistical Association*, 57 (June 1962), 269–326.

[2] ———, "The Anomalous Concept of Statistical Evidence," mimeographed technical report, IMM-NYU 332, Courant Inst. of Math. Sci., New York University, 1964.

[3] ———, "Likelihood," *International Encyclopedia of the Social Sciences*, New York: Macmillan and Free Press,1968.

[4] ———, "Concepts of Statistical Evidence," in Sidney Morgenbesser, Patrick Suppes, and Morton White, eds., *Essays in Honor of Ernest Nagel*, New York: St. Martin's Press, 1969.

[5] ———, "On Durbin's Modified Principle of Conditionality," *Journal of the American Statistical Association*, 65 (June 1970a), 402–3.

[6] ———, "Statistical Methods in Scientific Inference," *Nature*, 225 (March 14, 1970b), 1033. (Annotated version, containing proof corrections not made in publication, available from author.)

[7] ———, "A Perspective for Strengthening Scholarship in Statistics," *The American Statistician*, 25, No. 3 (June 1971), 14–7.

[8] ———, "The Random Phenotype Concept, with Applications," *Genetics*, 1972, (in press).

[9] Joshi, V.M., "A Note on Birnbaum's Theorem Relating to Conditionality, Sufficiency and Likelihood," typed manuscript, 1970, 1–11; Revised version, 1971, 1–7.

[10] Lindgren, B.W., *Statistical Theory*, New York: The Macmillan Co., 1960, 1962.

[11] Pratt, John W., "Review of E. Lehmann's *Testing Statistical Hypotheses*," *Journal of the American Statistical Association* 56 (March 1961), 163–7.

[12] ———, Comments on A. Birnbaum's "On the Foundations of Statistical Inference," *Journal of the American Statistical Association* 57 (June 1962), 314–5.

[13] Schwartz, J., "The Pernicious Influence of Mathematics on Science," in E. Nagel, P. Suppes, and A. Tarski, eds., *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, Stanford University Press, 1962.

[14] Tukey, John W., "The Future of Data Analysis," *Annals of Mathematical Statistics* 33 (March 1962), 1–67.