On the Foundations of Statistical Inference
Author(s): Allan Birnbaum
Source: *Journal of the American Statistical Association*, Vol. 57, No. 298 (Jun., 1962), pp. 269-306
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2281640
Accessed: 21/09/2009 23:23

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 298 JUNE, 1962 *Volume 57*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE[1]

ALLAN BIRNBAUM
New York University

The concept of conditional experimental frames of reference has a significance for the general theory of statistical inference which has been emphasized by R. A. Fisher, D. R. Cox, J. W. Tukey, and others. This concept is formulated as a *principle of conditionality,* from which some general consequences are deduced mathematically. These include the *likelihood principle,* which has not hitherto been very widely accepted, in contrast with the conditionality concept which many statisticians are inclined to accept for purposes of "informative inference." The likelihood principle states that the "evidential meaning" of experimental results is characterized fully by the likelihood function, without other reference to the structure of an experiment, in contrast with standard methods in which significance and confidence levels are based on the complete experimental model. The principal writers supporting the likelihood principle have been Fisher and G. A. Barnard, in addition to Bayesian writers for whom it represents the "directly empirical" part of their standpoint. The likelihood principle suggests certain systematic reinterpretations and revisions of standard methods, including "intrinsic significance and confidence levels" and "intrinsic standard errors," which are developed and illustrated. The close relations between non-Bayesian likelihood methods and Bayesian methods are discussed.

1. INTRODUCTION, SUMMARY, AND GENERAL CONCLUSIONS

THIS paper treats a traditional and basic problem-area of statistical theory, which we shall call *informative inference*, which has been a source of continuing interest and disagreement. The subject-matter of interest here may be called *experimental evidence:* when an experimental situation is represented by an adequate mathematical statistical model, denoted by E, and when any specified outcome x of E has been observed, then (E, x) is an instance of *statistical evidence*, that is, a mathematical model of an instance of experimental evidence. Part of the specification of E is a description of the range of unknown parameter values or of statistical hypotheses under consideration, that is, the description of a parameter space Ω of parameter points θ. The remaining part of E is given by description of the sample space of possible outcomes x of E, and of their re-

[1] This paper was presented at a special discussion meeting of the American Statistical Association on Wednesday, December 27, 1961 in the Roosevelt Hotel, New York City. George E. P. Box presided. Preprints of the paper were available several weeks before the meeting. Discussion which followed the presentation is reported on pp. 307 to 326 of this issue. Research on which the paper is based was supported by the Office of Naval Research.

spective probabilities or densities under respective hypotheses, typically by use of a specified probability density function $f(x, \theta)$ for each θ.

Methods such as significance tests and interval estimates are in wide standard use for the purposes of reporting and interpreting the essential features of statistical evidence. Various approaches to statistical theory have been concerned to an appreciable extent with this function. These include: Bayesian approaches, including those utilizing the principle of insufficient reason; some approaches using confidence methods of estimation and related tests of hypotheses; the fiducial approach of R. A. Fisher; and approaches centering on the direct inspection and interpretation of the likelihood function alone, as suggested by Fisher and G. A. Barnard. However the basic concepts underlying this function seem in need of further clarification.

We may distinguish two main general problems of informative inference: The problem of finding an appropriate *mathematical characterization* of statistical evidence as such; and the problem of *evidential interpretation,* that is, of determining concepts and terms appropriate to describe and interpret the essential properties of statistical evidence. It is useful sometimes to think of these problems, especially the first one, in connection with the specific function of reporting experimental results in journals of the empirical sciences.

The present analysis of the first problem begins with the introduction of the symbol $\text{Ev}(E, x)$ to denote the *evidential meaning* of a specified instance (E, x) of statistical evidence; that is, $\text{Ev}(E, x)$ stands for the essential properties (which remain to be clarified) of the statistical evidence, as such, provided by the observed outcome x of the specified experiment E. The next steps involve consideration of conditions under which we may recognize and assert that two instances of statistical evidence, (E, x) and (E', y), are equivalent in all relevant respects; such an assertion of *evidential equivalence* between (E, x) and (E', y) is written: $\text{Ev}(E, x) = \text{Ev}(E', y)$.

A first condition for such equivalence, which is proposed as an *axiom,* is related to the concept of sufficient statistic which plays a basic technical role in each approach to statistical theory. This is:

> *The principle of sufficiency* (S): If E is a specified experiment, with outcomes x; if $t = t(x)$ is any sufficient statistic; and if E' is the experiment, derived from E, in which any outcome x of E is represented only by the corresponding value $t = t(x)$ of the sufficient statistic; then for each x, $\text{Ev}(E, x) = \text{Ev}(E', t)$, where $t = t(x)$.

A familiar illustration of the concept formulated here is given by the problem of determining confidence limits for a binomial parameter: It is well known that exact confidence levels in this problem are achieved only with use of an auxiliary randomization variable, and that such confidence limits cannot be represented as functions of only the binomial sufficient statistic; the reluctance or refusal of many statisticians to use such confidence limits for typical purposes of informative inference is evidently an expression, within the context of this approach, of the principle formulated above. (S) may be described informally as asserting the "irrelevance of observations independent of a sufficient statistic."

A second condition for equivalence of evidential meaning is related to concepts of conditional experimental frames of reference; such concepts have been suggested as appropriate for purposes of informative inference by writers of several theoretical standpoints, including Fisher and D. R. Cox. This condition concerns any experiment E which is mathematically equivalent to a *mixture* of several other *component* experiments E_h, in the sense that observing an outcome x of E is mathematically equivalent to observing first the value h of random variable having a known distribution (not depending upon unknown parameter values), and then taking an observation x_h from the component experiment E_h labeled by h. Then (h, x_h) or (E_h, x_h) is an alternative representation of the outcome x of E. The second proposed axiom, which many statisticians are inclined to accept for purposes of informative inference, is:

> *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 $\mathrm{Ev}(E, (E_h, x_h)) = \mathrm{Ev}(E_h, x_h)$. That is, the evidential meaning of any outcome of any mixture experiment is the same as that of the corresponding outcome of the corresponding component experiment, ignoring the over-all structure of the mixture experiment.

(C) may be described informally as asserting the "irrelevance of (component) experiments not actually performed."

The next step in the present analysis concerns a third condition for equivalence of evidential meaning, which has been proposed and supported as self-evident principally by Fisher and by G. A. Barnard, but which has not hitherto been very generally accepted. This condition concerns the likelihood function, that is, the function of θ, $f(x, \theta)$, determined by an observed outcome x of a specified experiment E; two likelihood functions, $f(x, \theta)$ and $g(y, \theta)$, are called the same if they are proportional, that is if there exists a positive constant c such that $f(x, \theta) = cg(y, \theta)$ for all θ. This condition is:

> *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 $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', y)$. That is, the evidential meaning of any outcome x of any experiment E is characterized fully by giving the likelihood function $cf(x, \theta)$ (which need be described only up to an arbitrary positive constant factor), without other reference to the structure of E.

(L) may be described informally as asserting the "irrelevance of outcomes not actually observed."

The fact that relatively few statisticians have accepted (L) as appropriate for purposes of informative inference, while many are inclined to accept (S) and (C), lends interest and significance to the result, proved herein, that (S) *and* (C) *together are mathematically equivalent to* (L). When (S) and (C) are adopted, their consequence (L) constitutes a significant solution to the first problem of informative inference, namely that a mathematical characterization of statistical evidence as such is given by the likelihood function.

For those who find (S) and (C) compellingly appropriate (as does the present writer), their consequence (L) has immediate radical consequences for the every-day practice as well as the theory of informative inference. One basic consequence is that reports of experimental results in scientific journals should in principle be descriptions of likelihood functions, when adequate mathematical-statistical models can be assumed, rather than reports of significance levels or interval estimates. Part II of this paper, Sections 6–13, is concerned with the general problem of evidential interpretation, on the basis of the likelihood principle.

(L) implies that experimental frames of reference, whether actual, conditional, or hypothetical, have no necessary essential role to play in evidential interpretations. But most current statistical practice utilizes concepts and techniques of evidential interpretation (like significance level, confidence interval, and standard error) based on experimental frames of reference. Hence it seems of considerable practical and heuristic value, as well as of theoretical interest, to consider how far the commonly used concepts and techniques can be reinterpreted or revised to provide modes of describing and interpreting likelihood functions as such, utilizing experimental frames of reference in a systematic but clearly conventional manner compatible with (L). This approach leads to concepts and techniques of evidential interpretation called "intrinsic significance levels," "intrinsic confidence sets, with intrinsic confidence levels," and "intrinsic standard error of an estimate"; these are illustrated by examples. Perhaps the principal value of this approach will be to facilitate understanding and use of likelihood functions as such, in the light of the likelihood principle, by relating them to concepts and techniques more familiar to many statisticians.

Bayesian methods based on the principle of insufficient reason, and a version of Fisher's fiducial argument, are interpreted as alternative partly-conventional modes of description and evidential interpretation of likelihood functions. Many points of formal coincidence between these and intrinsic confidence methods are noted.

This analysis shows that when informative inference is recognized as a distinct problem-area of mathematical statistics, it is seen to have a scope including some of the problems, techniques, and applications customarily subsumed in the problem-areas of point or set estimation, testing hypotheses, and multidecision procedures. In fact the course of development of the latter areas of statistics seems to have been shaped appreciably by the practice of formulating problems of informative inference as problems of one of these kinds, and developing techniques and concepts in these areas which will serve adequately for informative inference. At the same time each of these methods can serve purposes distinct from informative inference; the inclusion of problems of two distinct kinds, one of them traditional but not clearly enough delineated, seems to have forced a certain awkwardness of formulation and development on these areas. For example, problems of estimation of a real-valued parameter have traditionally been dealt with by techniques which supply a point estimate supplemented by an index of precision, or an interval estimate, and such techniques serve the purposes of informative inference fairly adequately, particularly in problems of simple structure. However, in modern generalizations and refinements of theories of estimation it becomes clear that no single formulation is appropriate in general to serve the distinct functions of informative inference on the one hand and either point or interval estimation on the other hand; and that the attempt to serve both functions by a single formal theory and set of techniques makes for awkwardness and indistinctness of purpose.

Recognition of informative inference as a distinct problem-area with its own basic concepts and appropriate techniques should help unburden the other problem-areas of statistics, particularly statistical decision theory, for freer developments more clearly and deeply focused on the problems in their natural mathematical and practical scope. Tukey [20, pp. 450, 468–74]; [21] has recently emphasized that the "elementary" problems of mathematical statistics are still with us as live problems. Among these must be included questions of specification of "what are the problems and the problem-areas of mathematical statistics, what is their formal mathematical and extra-mathematical content, and what are their scopes of application?" For example, what are the typical substantial functions of point and interval estimation, and of tests of hypotheses, apart from the function of informative inference?

The fact that the likelihood principle follows from the principles of sufficiency and conditionality, which many find more acceptable than Bayes' principle, seems to provide both some comfort and some challenge to Bayesian viewpoints: The "directly empirical" part of the Bayesian position concerning the role of the likelihood function is given new support independent of Bayes' principle itself. But this suggests the question: What are the specific contributions of the Bayesian concepts and techniques to the interpretation and use of statistical evidence, above and beyond what is possible by less formalized interpretations and applications based on direct consideration of the likelihood function in the light of other aspects of the inference situation, without formal use of prior probabilities and Bayes' formula? Specifically, what are the precise contributions of quantitative prior probabilities, and of the other formal parts of the Bayesian methods? Evidently in the present state of our understanding there can be interesting collaboration between Bayesian and non-Bayesian statisticians, in exploring the possibilities and limitations of both formal and informal modes of interpreting likelihood functions, and in developing the important problem-areas of experimental design and of robustness from the standpoint of such interpretations.

These considerations also present some challenge to non-Bayesian statisticians accustomed to use of standard techniques of testing and estimation, in which error-probabilities appear as basic terms of evidential interpretation in a way which is incompatible with the principle of conditionality. The writer has not found any apparent objections to the latter principle which do not seem to stem from notions of "conditional" distinct from that considered here, or else from purposes other than the modest but important one of informative inference.

Part I

2. *Statistical evidence.* A traditional standard in empirical scientific work is accurate reporting of "what was observed, and under what experimental plan and conditions." Such reports are an essential part of the literature and the

structure of the empirical sciences; they constitute the body of observational or *experimental evidence* available at any stage to support the practical applications and the general laws, theories, and hypotheses of the natural sciences. (Cf. Wilson [25], especially Ch. 13, and references therein.)

In some circumstances the "experimental plan and conditions" can be represented adequately by a mathematical-statistical model of the experimental situation. The adequacy of any such model is typically supported, more or less adequately, by a complex informal synthesis of previous experimental evidence of various kinds and theoretical considerations concerning both subject-matter and experimental techniques. (The essential place of working "conclusions" in the fabric and process of science has been discussed recently by Tukey [22].) We deliberately delimit and *idealize* the present discussion by considering only models whose adequacy is postulated and is not in question.

Let E denote a mathematical-statistical model of a given experimental situation: When questions of experimental design (including choice of sample size or possibly a sequential sampling rule) have been dealt with, the sample space of possible outcomes x of E is a specified set $S = \{x\}$. We assume that each of the possible distributions of X is labeled by a parameter point θ in a specified parameter space $\Omega = \{\theta\}$, and is represented by a specified elementary probability function $f(x, \theta)$. The probability that E yields an outcome x in A is

$$P(A \mid \theta) \equiv \text{Prob}(X \in A \mid \theta) = \int_A f(x, \theta) d\mu(x),$$

where μ is a specified (σ-finite) measure on S, and A is any (measurable) set. Thus any mathematical model of an experiment, E, is given by specifying its mathematical ingredients: (Ω, S, f, μ). (No methods of advanced probability theory are used in this paper. The reader familiar only with probabilities defined by

$$P(A \mid \theta) = \sum_{x \in A} f(x, \theta),$$

for discrete distributions, and by

$$P(A \mid \theta) = \int_A f(x, \theta) dx,$$

for continuous distributions (with dx possibly representing $dx_1 \cdots dx_n$), can regard the symbol $\int_A f(x, \theta) d\mu(x)$ as a generalization including those two important cases and some others.)

In an experimental situation represented by such a model E, the symbol (E, x) denotes an instance of *statistical evidence*. The latter term will be used here to denote any such mathematical model of an instance of experimental evidence: x represents "what was observed," and E represents "under what experimental plan and conditions."

The central purpose of this paper is to clarify the essential structure and properties of statistical evidence in various instances. We use the symbol $\text{Ev}(E, x)$, and the term *evidential meaning* (of a specified outcome x of a specified experiment E), to refer to these essential properties and structure, whose precise nature remains to be discussed.

The first general problem to be considered (throughout Part I) is whether a satisfactory *mathematical characterization* can be found for evidential meaning in various instances. The second general purpose (in the following sections, Part II) is to consider what concepts, terms, and techniques are appropriate for representing, interpreting, and expressing evidential meaning in various instances; in other words, to consider critically the function of *evidential interpretation* of experimental results. The broad but delimited part of mathematical statistics which is concerned with these two problems, the characterization and the interpretation of statistical evidence as such, will be termed here the problem-area of *informative* (statistical) *inference.* While such problems and methods have broad and varied relevance and use, it will be helpful sometimes to focus attention on the specific and relatively simple function referred to above: the formal reporting of experimental results, in empirical scientific journals, in terms which are appropriate to represent their character as evidence relevant to parameter values or statistical hypotheses of interest. We restrict present consideration to situations in which all questions of characterizing and interpreting statistical evidence will have been considered in full generality before an experiment is carried out: Our discussion concerns all possible outcomes x and possible interpretations thereof, as these can in principle be considered at the outset of a specified experiment; such discussion can subsequently be broadened to include questions of appraisal, comparison, and design of experiments for purposes of informative inference. Our discussion will not touch on tests or other modes of inference in cases where the set of possible alternative distributions is not specified initially [9, Ch. 3].

Since the problem-area of informative inference has not received a generally accepted delineation or terminology, it will be useful to note here some of the terms and concepts used by writers representing several different approaches:

a) R. A. Fisher [9, pp. 139–41] has employed the term "estimation" to refer to this problem-area, in contrast with the widely current usage of this term to refer to problems of interval (or set) or point estimation. Fisher's paper [10, pp. 175–6] includes in its introductory section ("On the nature of the problem") the following interpretation of Gossett's fundamental work on testing a normal mean: "In putting forth his test of significance 'Student' (1908) specified that the problem with which he is concerned is that of a *unique* sample. His clear intention in this is to exclude from his discussion all possible suppositions as to the 'true' distribution of the variances of the populations which might have been sampled. If such a distribution were supposed known, 'Student's' method would be open to criticism and to correction. In following his example it is not necessary to deny the existence of knowledge based on previous experience, which might modify his result. It is sufficient that we shall deliberately choose to examine the evidence of the sample on its own merits only."

The last two sentences may be taken to be descriptive of the problem-area of informative inference, even though the context refers to significance tests. It is clear that many of the principal modern statistical concepts and methods developed by Fisher and other non-Bayesian writers have been directed to problems of informative inference. This applies in particular to Fisher's description of three modes of statistical inference, significance tests, estimation (in the broad sense indicated above), and the fiducial argument [9, Ch. 3, especially p. 73].

While such phrases as "specification of uncertainty" and "measure of the rational grounds for . . . disbelief" have sometimes been used [9, pp. 43–4] to describe the purpose and nature of informative inference, it is possible and it seems desirable to discuss these problems without use of terms having specifically subjective or psychological reference. The latter course will be followed throughout the present paper; our discussion of the structure and properties of statistical evidence will not involve terms or concepts referring to "reactions to evidence" in any sense.

b) Many of the developments and applications of statistical methods of testing and estimation which stem from the work of Neyman and Pearson have been directed to informative inference. Such methods are widely considered to serve this purpose fairly adequately and soundly. The basic terms of such applications and interpretations are probabilities of the errors of various kinds which could be made in connection with a given experiment. (Measures of precision of estimators can be interpreted as referring to probabilities of various possible errors in estimation.) It is considered an essential feature of such interpretations that these basic error-probability terms are objective, in the mathematical sense (and in the related physical sense) that conceptually-possible repetitions of an experiment, under respective hypotheses, would generate corresponding relative frequencies of errors. In typical current practice, some reference to such error-probabilities accompanies inference statements ("assertions," or "conclusions") about parameter values or hypotheses. If an inference is thus accompanied by relevant error-probabilities which are fairly small, the inference is considered supported by fairly strong evidence; if such relevant error-probabilities are all very small, the evidence is considered very strong. These remarks simply describe the general nature of evidential interpretations of experimental results, which is traditionally and widely recognized in scientific work; here the concepts and techniques of testing and estimation serve as frameworks for such evidential interpretations of results. Such evidential interpretations do not seem to differ in kind from those associated with the less technical notion of circumstantial evidence when all relevant hypotheses are considered (cf. for example Cohen & Nagel [7], pp. 347–51); they differ sharply in degree, in that precisely specified frameworks for such interpretations are provided by the mathematical models of experiments and by the formal definitions and properties of the inference methods employed.

The usefulness for informative inference of tests and especially of confidence set estimates has been emphasized recently by several writers, including Cox [8], Tukey [22], and Wallace [23], [24]. At the same time these writers have been concerned also with technical and conceptual problems related to such use and interpretation of these methods. Cox [8, p. 359] has cited the term "summarization of evidence" to indicate the function of informative inference, and like some other writers has described it as concerned with "statistical inferences" or "conclusions," in contrast with statistical decision problems for which the basic mathematical structure and interpretations seem relatively clear. As Cox writes [8, p. 354], "it might be argued that in making an inference we are 'deciding' to make a statement of a certain type about the populations and that, therefore, provided the word decision is not interpreted too narrowly, the study of statistical decisions embraces that of inferences. The point here is that one of the main general problems of statistical inference consists in deciding what types of statement can usefully be made and exactly what they mean. In statistical decision theory, on the other hand, the possible decisions are considered as already specified."

c) Approaches to statistical inference problems based upon Bayes' principle of inverse probability (with any interpretation) obtain on that basis clear and simple answers to questions of informative inference, as will be reviewed below. Writing from his own Bayesian standpoint, Savage [18] has recently described as follows the difficulties and prospects of non-Bayesian approaches such as those discussed above: "Rejecting both necessary and personalistic views of probability left statisticians no choice but to work as best they could with frequentist views. . . . The frequentist is required, therefore, to seek a concept of evidence, and of reaction to evidence, different from that of the primitive, or natural, concept that is tantamount to application of Bayes' theorem.

"Statistical theory has been dominated by the problem thus created, and its most profound and ingenious efforts have gone into the search for new meanings for the concepts of inductive inference and inductive behavior. Other parts of this lecture will at least suggest concretely how these efforts have failed, or come to a stalemate. For the moment, suffice it to say that a problem which after so many years still resists solution is suspect of being ill formulated, especially since this is a problem of conceptualization, not a technical mathematical problem like Fermat's last theorem or the four-color problem."

The present paper is concerned primarily with approaches to informative inference which do not depend upon the Bayesian principle of inverse probability.

3. *The principle of sufficiency.* As the first step of our formal analysis of the structure of evidential meaning, $\mathrm{Ev}(E, x)$, we observe that certain cases of *equivalence of evidential meaning* can be recognized, even in advance of more explicit characterization of the nature of evidential meaning itself. We shall write $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', y)$ to denote that two instances of statistical evidence, (E, x) and (E', y), have the same (or equivalent) evidential meaning.

For example, let (E, x) and (E', y) be any two instances of statistical evidence, with E and E' having possibly different mathematical structures but the same parameter space $\Omega = \{\theta\}$. Suppose that there exists a one-to-one transformation of the sample space of E onto the sample space of E': $y = y(x)$, $x = x(y)$, such that the probabilities of all corresponding (measurable) sets under all corresponding hypotheses are equal: $\mathrm{Prob}(Y \in A' \mid \theta) = \mathrm{Prob}(X \in A \mid \theta)$ if

$A' = y(A)$. Then the models E and E' are *mathematically equivalent*, one being a relabeling of the other. If respective outcomes x of E and y of E' are related by $y = y(x)$, they also are mathematically equivalent, and the two instances of statistical evidence (E, x) and (E', y) may be said to have the same evidential meaning: $\text{Ev}(E, x) = \text{Ev}(E', y)$. A simple concrete example is that of models of experiments which differ only in the units in which measurements are expressed.

Again, consider (E, x) and (E', t), where $t(x)$ is any *sufficient* statistic for E, and where E' represents the possible distributions of $t(x)$ under the respective hypotheses of E. Then, for reasons which are recognized within each approach to statistical theory, we may say that $\text{Ev}(E, x) = \text{Ev}(E', t)$ if $t = t(x)$. An example which occurs within the approach to informative inference which utilizes confidence intervals (and related tests) involves the possible use of randomized confidence limits (or tests), for example for a binomial parameter. The view, held by many, that randomized forms of such techniques should not be used seems to stem from an appreciation that sufficiency concepts must play a certain guiding role in the development of methods appropriate for informative inference. (For a recent discussion and references, cf. [21].)

Such considerations may be formalized as follows to provide an *axiom* which we adopt to begin our mathematical characterization of evidential meaning:

> *Principle of sufficiency* (S): Let E be any experiment, with sample space $\{x\}$, and let $t(x)$ be any sufficient statistic (not necessarily real-valued). Let E' denote the derived experiment, having the same parameter space, such that when any outcome x of E is observed the corresponding outcome $t = t(x)$ of E' is observed. Then for each x, $\text{Ev}(E, x) = \text{Ev}(E', t)$, where $t = t(x)$.

It is convenient to note here for later use certain definitions and a mathematical consequence of (S): If x is any specified outcome of any specified experiment E, the *likelihood function determined by* x is the function of θ: $cf(x, \theta)$, where c is assigned arbitrarily any positive constant value. Let E and E' denote any two experiments with the same parameter space (E' could be identical with E), and let x and y be any specified outcomes of these respective experiments, determining respective likelihood functions $f(x, \theta)$ and $g(y, \theta)$; if for some positive constant c we have $f(x, \theta) = cg(y, \theta)$ for all θ, x and y are said to determine *the same likelihood function*. It has been shown in the general theory of sufficient statistics (cf. [1]) that if two outcomes x, x' of *one* experiment E determine the same likelihood function (that is, if for some positive c we have $f(x, \theta) = cf(x', \theta)$ for all θ), then there exists a (minimal) sufficient statistic t such that $t(x) = t(x')$. (In the case of any discrete sample space, the proof is elementary.) This, together with (S), immediately implies

Lemma 1. If two outcomes x, x' of any experiment E determine the same likelihood function, then they have the same evidential meaning: $\text{Ev}(E, x) = \text{Ev}\ (E, x')$.

4. *The principle of conditionality*

4.1. The next step in our analysis is the formulation of another condition for equivalence of evidential meaning, which concerns conditional experimental frames of reference. This will be stated in terms of the following definitions:

An experiment E is called a *mixture* (or a mixture experiment), with *components* $\{E_h\}$, if it is mathematically equivalent (under relabeling of sample points) to a two-stage experiment of the following form:

(a) An observation h is taken on a random variable H having a fixed and known distribution G. (G does not depend on unknown parameter values.)

(b) The corresponding component experiment E_h is carried out, yielding an outcome x_h.

Thus each outcome of E is (mathematically equivalent to) a pair (E_h, x_h). (Each component experiment E_h, and E, all have the same parameter space. Every experiment is a mixture in the trivial sense that all components may be identical; the non-trivial cases, with non-equivalent components, are of principal interest. Examples will be discussed below.)

As a second proposed *axiom* concerning evidential meaning, we take the

> *Principle of conditionality* (C)*:* If an experiment E is (mathematically equivalent to) a mixture G of components $\{E_h\}$, with possible outcomes (E_h, x_h), then
>
> $$\mathrm{Ev}(E, (E_h, x_h)) = \mathrm{Ev}(E_h, x_h).$$
>
> That is, the evidential meaning of any outcome (E_h, x_h) of any experiment E having a mixture structure is the same as: the evidential meaning of the corresponding outcome x_h of the corresponding component experiment E_h, ignoring otherwise the over-all structure of the original experiment E.

4.2. A number of writers have emphasized the significance of conditionality concepts for the analysis of problems of informative inference. Fisher recently wrote [9, pp. 157–8] "The most important step which has been taken so far to complete the structure of the theory of estimation is the recognition of Ancillary statistics." (Evidently a statistic like h above, whose distribution is known and independent of unknown parameters, is an example of an ancillary statistic. "Estimation" is used here by Fisher in the broad sense of informative inference, rather than point or interval estimation.) Other relevant discussions have been given by Cox [8, pp. 359–63], Wallace [23, especially p. 864 and references therein], and Lehmann [14, pp. 139–40].

The following sections will be largely devoted to the deduction of some mathematical consequences of (C) and (S), and to their interpretation. The remainder of the present section is devoted to discussion and illustration of the meaning of (C); and to illustration of the considerations which seem to many statisticians, including the writer, to give compelling support to adoption of (C) as an appropriate extra-mathematical assertion concerning the structure of evidential meaning.

It can be shown that (S) is implied mathematically by (C). (The method of proof is the device of interpreting the conditional distribution of x, given $t(x)=t$, as a distribution $G_t(h)$ defining a mixture experiment equivalent to the given experiment.) This relation will not be discussed further here, since there seems to be little question as to the appropriateness of (S) in any case.

4.3. *Example.* A simple concrete (but partly hypothetical) example is the following: Suppose that two instruments are available for use in an experiment whose primary purpose is informative inference, for example, to make observations on some material of general interest, and to report the experimental results in appropriate terms. Suppose that the experimental conditions are fixed, and that these entail that the selection of the instrument to be used depends upon chance factors not related to the subject-matter of the experiment, in such a way that the instruments have respective known probabilities $g_1 = .73$ and $g_2 = .27$ of being selected for use. The experimental conditions allow use of the selected instrument to make just one observation, and each instrument gives only dichotomous observations, $y = 1$ ("positive") or 0 ("negative").

(We recall that discussion of design of experiments for informative inference has been deferred; but we stress that any satisfactory general analysis of evidential meaning must deal adequately with artificial and hypothetical experiments as well as with those of commonly-encountered forms. Even the present example is not very artificial, since the alternative instruments are simple analogues of observable experimental conditions (like independent variables in some regression problems) which may be uncontrollable and which have known effects on experimental precision.) If the instruments are labeled by $h = 1$ or 2, respectively, then each outcome of this experiment E is represented by a symbol (h, y) or (h, y_h), where $h = 1$ or 2, and $y = y_h = 0$ or 1. We assume that the material under investigation is known to be in one of just two possible states, H_1 or H_2 (two simple hypotheses). Each instrument has equal probabilities of "false positives" and of "false negatives." For the first instrument these are

$$\alpha_1 = \text{Prob}(Y_1 = 1 \mid H_1) = \text{Prob}(Y_1 = 0 \mid H_2) = \frac{1}{730} \doteq .0014,$$

and for the second instrument

$$\alpha_2 = \text{Prob}(Y_2 = 1 \mid H_1) = \text{Prob}(Y_2 = 0 \mid H_2) = .10.$$

As an instance of the general proposition (C), consider the assertion: $\text{Ev}(E, (E_1, 1)) = \text{Ev}(E_1, 1)$. This assertion is apparently not necessary on mathematical grounds alone, but it seems to be supported compellingly by considerations like the following concerning the nature of evidential meaning: Granting the validity of the model E and accepting the experimental conditions which it represents, suppose that E leads to selection of the first instrument (that is, $H = h = 1$ is observed). Then by good fortune the experimenter finds himself in the same position as if he had been assured use of that superior instrument (for one observation) as an initial condition of his experiment. In the latter hypothetical situation, he would be prepared to report either $(E_1, 0)$ or $(E_1, 1)$ as a complete description of the statistical evidence obtained. In the former actual situation, the fact that the first instrument might not have been selected seems not only hypothetical but completely irrelevant: For purposes of informative inference, if $Y = 1$ is observed with the first instrument, then the report $(E_1, 1)$ seems to be an appropriate and complete description of the statistical evidence obtained; and the "more complete" report $(E, (E_1, 1))$ seems to differ from it only by the addition of recognizably redundant elements irrelevant to the evidential meaning and evidential interpretation of this outcome of E. The latter redundant elements are the descriptions of other component experiments (and their probabilities) which might have been carried out but in fact were not. Parallel comments apply to the other possible outcomes of E.

4.4. As formulated above, (C) is not a recommendation (or directive or convention) to replace unconditional by conditional experimental frames of reference wherever (C) is seen to be applicable. However if (C) is adopted it tends to invite such application, if only for the advantage of parsimony, since a conditional frame of reference is typically simpler and seems more appropriately refined for purposes of informative inference. Writers who have seen value in such conditionality concepts have usually focused attention on their use in this way. However, even the range of such applications has not been fully investigated in experiments of various structures. And the implications of such conditionality concepts for problems of informative inference in general appear considerably more radical than has been generally anticipated, as will be indicated below. We shall be primarily concerned with the use of (C) as a tool in the formal analysis of the structure of evidential meaning; and in such use, (C) as formulated above also sanctions the replacement of a conditional experimental frame of reference by an appropriately corresponding unconditional one (by substitution of $\mathrm{Ev}(E, (E_h, x_h))$ for an equivalent $\mathrm{Ev}(E_h, x_h)$).

4.5. Another aspect of such interpretations can be discussed conveniently in terms of the preceding example. The example concerned an experiment whose component experiments are based on one or another actual experimental instrument. Consider next an alternative experiment plan (of a more familiar type) which could be adopted for the same experimental purpose: Here just one instrument is available, the second one described above, which gives observations $Y=1$ with probabilities .1 and .9 under the same respective hypotheses H_1, H_2, and otherwise gives $Y=0$. The present experimental plan, denoted by E_B, calls for 3 independent observations by this instrument; thus the model E_B is represented by the simple binomial distributions of

$$X = \sum_{j=1}^{3} Y_j:$$

$$H_1: f_1(x) = \binom{3}{x}(.1)^x(.9)^{3-x},$$

$$H_2: f_2(x) = \binom{3}{x}(.9)^x(.1)^{3-x},$$

for $x=0$, 1, 2, 3. E_B will provide one of the instances of statistical evidence (E_B, x), $x=0$, 1, 2, or 3. The *physical* experimental procedures represented respectively by E and E_B are manifestly different. But we verify as follows that the mathematical-statistical models E and E_B are *mathematically* equivalent: Each experiment leads to one of four possible outcomes, which can be set in the following one-to-one correspondence:

E yields: $(E_h, y_h) =$	$(E_1, 0)$	$(E_2, 0)$	$(E_2, 1)$	$(E_1, 1)$
E_B yields: $x =$	0	1	2	3

It is readily verified that under each hypothesis the two models specify identical probabilities for corresponding outcomes. For example,

$$\begin{aligned}\text{Prob}((E_1, 0) \mid H_1, E) &= (.73)\frac{729}{730} = .729\\ &= \binom{3}{0}(.1)^0(.9)^3 = f_1(0)\\ &= \text{Prob}(X = 0 \mid H_1, E_B).\end{aligned}$$

Thus $(E_B, 0)$ and $(E, (E_1, 0))$ are *mathematically equivalent* instances of statistical evidence. We therefore write $\text{Ev}(E_B, 0) = \text{Ev}(E, (E_1, 0))$. Is the latter assertion of equivalence of evidential meanings tenable here, because of the *mathematical* equivalence of $(E_B, 0)$ and $(E, (E_1, 0))$ alone, and despite the gross difference of *physical* structures of the experiments represented by E_B and E? An affirmative answer seems necessary on the following *formal* grounds: Each of the models E and E_B was assumed to be an adequate mathematical-statistical model of a corresponding physical experimental situation; this very strong assumption implies that there are no physical aspects of either situation which are relevant to the experimental purposes except those represented in the respective models E and E_B. The latter models may be said to represent adequately and completely the assumed *physical* as well as mathematical structures of the experiments in all *relevant* respects; for example, the usual conceptual frequency interpretations of all probability terms appearing in each model may be taken to characterize fully the physical structure and meaning of each model. Hence the assumed adequacy and the mathematical equivalence of the two models imply that the two experimental situations have in effect been assumed to be physically equivalent in all *relevant* respects. This interpretative conclusion can be illustrated further by considering the rhetorical question: On what theoretical or practical grounds can an experimenter reasonably support any definite preference between the experiments represented by E and E_B, for *any* purpose of statistical inference or decision-making, assuming the adequacy of each model?

Combining this discussion with section 4.3 above, we find that (C) implies that $\text{Ev}(E_B, 0) = \text{Ev}(E_1, 0)$, although no mixture structure was apparent in the physical situation represented by E_B, nor in the binomial model E_B as usually interpreted.

4.6. We note that (C) above differs in meaning and scope from the purely technical use which is sometimes made of conditional experimental frames of reference, as in the development of similar tests of composite hypotheses (as in Lehmann [14, p. 136]) or of best unbiased estimators.

4.7. We note also that (C) above does not directly involve, or ascribe meaning to, any notion of evidential interpretations "conditional on an observed sample point x." Rather, (C) ascribes equivalence to certain instances of evidential meaning of respective outcomes, each referred to a specified mathematically complete experimental frame of reference. (The phrase in quotes can be given a precise mathematical meaning under postulation of the principle of inverse probability, in which case it refers to a posterior distribution, given x. However our discussion is not based on such postulation.)

4.8. In considering whether (C) seems appropriate for all purposes of informative inference, it is necessary to avoid confusion with still another usage of "conditional" which differs from that in (C). A familiar simple example of this other usage occurs in connection with a one-way analysis of variance experiment under common normality assumptions. Results of such an experiment may be interpreted either "conditionally" (Model I) or "unconditionally" (Model II), and in some situations there are familiar purposes of informative inference (focusing on a component of variance) in which the "unconditional" interpretation is useful and necessary. However, the latter important point is *not* relevant to the question of the general appropriateness of (C) for informative inference, because the "conditional" frame of reference in this example cannot be interpreted as a component experiment within a mixture experiment as required for applicability of (C).

4.9. It is the opinion of the writer (among others) that upon suitable consideration the principle of conditionality will be generally accepted as appropriate for purposes of informative inference, and that apparent reservations will be found to stem either from purposes which can usefully be distinguished from informative inference, or from interpretations of "conditionality" different from that formulated in (C), some of which have been described above. (Of course purposes of several kinds are frequently represented in one experimental situation, and these are often served best by applying different concepts and techniques side by side as appropriate for the various purposes.) In any case, the following sections are largely devoted to examination of the *mathematical* consequences of (C) and their interpretation.

5. *The likelihood principle*

5.1. The next step in our analysis concerns a third condition for equivalence of evidential meaning:

> *The likelihood principle* (L)*:* If E and E' are any two experiments with a common parameter space, and if x and y are any respective outcomes which determine likelihood functions satisfying $f(x, \theta) = cg(y, \theta)$ for some positive constant $c = c(x, y)$ and all θ, then $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', y)$. That is, the evidential meaning $\mathrm{Ev}(E, x)$ of any outcome x of any experiment E is characterized completely by the likelihood function $cf(x, \theta)$, and is otherwise independent of the structure of (E, x).

5.2. (L) is an immediate consequence of Bayes' principle, when the latter (with any interpretation) is adopted. Our primary interest, as mentioned, is in approaches which are independent of this principle.

5.3. Fisher [9, pp. 68–73, 128–31, and earlier writings] and Barnard [2, and earlier writings] have been the principal authors supporting the likelihood principle on grounds independent of Bayes' principle. (The principle of maximum likelihood, which is directed to the problem of point-estimation, is not to be identified with the likelihood principle. Some connections between the distinct problems of point-estimation and informative inference are discussed below.) Self-evidence seems to be essential ground on which these writers support (L).

5.4. Other modes of support for (L), such as the basic technical role of the likelihood function in the theory of sufficient statistics and in the characterization of admissible statistical decision functions, seem heuristic and incomplete, since (as in the formulation of (S), and its consequence Lemma 1, in Section 3 above) they do not demonstrate that evidential meaning is independent of the structure of an experiment apart from the likelihood function.

5.5. Far fewer writers seem to have found (L) as clearly appropriate, as an extra-mathematical statement about evidential meaning, as (C). It is this fact which seems to lend interest to the following:

Lemma 2. (L) implies, and is implied by, (S) and (C).

Proof: That (L) implies (C) follows immediately from the fact that in all cases the likelihood functions determined respectively by $(E, (E_h, x_h))$ and (E_h, x_h) are proportional. That (L) implies (S) follows immediately from Lemma 1 of Section 3.

The relation of principal interest, that (S) and (C) imply (L), is proved as follows: Let E and E' denote any two (mathematical models of) experiments, having the common parameter space $\Omega = \{\theta\}$, and represented by probability density functions $f(x, \theta)$, $g(y, \theta)$ on their respective sample spaces $S = \{x\}$, $S' = \{y\}$. (S and S' are to be regarded as distinct, disjoint spaces.) Consider the (hypothetical) mixture experiment E^* whose components are just E and E', taken with equal probabilities. Let z denote the generic sample point of E^*, and let C denote any set of points z; then $C = A \cup B$, where $A \subset S$ and $B \subset S'$, and

$$\begin{aligned}\text{Prob}(Z \in C \mid \theta) &= \frac{1}{2}\text{Prob}(A \mid \theta, E) + \frac{1}{2}\text{Prob}(B \mid \theta, E') \\ &= \frac{1}{2}\int_A f(x, \theta) d\mu(x) + \frac{1}{2}\int_B g(y, \theta) d\nu(y)\end{aligned}$$

(where A and B are measurable sets). Thus the probability density function representing E^* may be denoted by

$$h(z, \theta) = \begin{cases} \frac{1}{2}f(x, \theta), & \text{if } z = x \in S, \\ \frac{1}{2}g(y, \theta), & \text{if } z = y \in S'. \end{cases}$$

Each outcome z of E^* has a representation

$$z = \begin{cases} (E, x), & \text{if } z = x \in S, \\ (E', y), & \text{if } z = y \in S'. \end{cases}$$

From (C), it follows that

$$\begin{aligned}\text{Ev}(E^*, (E, x)) &= \text{Ev}(E, x), &&\text{for each } x \in S, \text{ and} \\ \text{Ev}(E^*, (E', y)) &= \text{Ev}(E', y), &&\text{for each } y \in S'.\end{aligned} \tag{5.1}$$

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. Then we have $h(x', \theta) \equiv ch(y', \theta)$ for all θ; that is, the two outcomes (E, x'), (E', y') of E^* determine the same likelihood function. Then it follows from (S) and its consequence Lemma 1 in Section 3 that

$$\text{Ev}(E^*, (E, x')) = \text{Ev}(E^*, (E', y')). \qquad (5.2)$$

From (5.1) and (5.2) it follows that

$$\text{Ev}(E, x') = \text{Ev}(E', y'). \qquad (5.3)$$

But (5.3) states that any two outcomes x', y' of any two experiments E, E' (with the same parameter space) have the same evidential meaning if they determine the same likelihood function. This completes the proof of equivalence of (L) with (S) and (C).

5.6. For those who adopt (C) and (S), their consequence (L) gives an explicit solution to our first general problem, the mathematical characterization of statistical evidence as such. The question whether different likelihood functions (on the same parameter space) represent different evidential meanings is given an affirmative answer in the following sections, in terms of evidential interpretations of likelihood functions on parameter spaces of limited generality; and presumably this conclusion can be supported quite generally.

5.7. The most important general consequence of (L) (and of (C)) for problems of *evidential interpretation* seems to be the following: Those modes of representing evidential meaning which include reference to any specific experimental frame of reference (including the actual one from which an outcome was obtained) are somewhat unsatisfactory; in particular, they tend to conceal equivalences between instances of evidential meaning which are recognizable under (L). Various modes of *interpretation* of evidential meaning will be discussed in the following sections, with particular attention to their relations to (L).

5.8. The scope of the role of ancillary statistics in informative inference seems altered in the light of the result that (C) and (S) imply (L). As mentioned, the usual use of (C) has depended on recognition of an ancillary statistic (or mixture structure) in the model of an actual experiment under consideration; and has consisted primarily of the adoption of conditional frames of reference, when thus recognized, for evidential interpretations. But the range of existence of ancillary statistics in experiments of various structures has not been completely explored; indeed, in the simple case of binary experiments (those with two-point parameter spaces), the fact that they exist in all but the simplest cases has been seen only very recently in reference [3]. Thus the potential scope and implications, which even such usual applications of (C) might have for informative inference, have not been fully seen.

Moreover, the question of conditions for uniqueness of ancillary statistics, when they exist, has received little attention. But simple examples have been found, some of which are described in reference [3], in which one experiment admits several essentially different ancillary statistics; when (C) is applied in the usual way to each of these alternative ancillary statistics in turn, one can obtain quite different conditional experimental frames of reference for evidential interpretation of a single outcome. Even isolated examples of this kind seem to pose a basic problem for this approach: it would seem that, in the face of such examples, the usual use of (C) must be supplemented either by a convention restricting its scope, or by a convention for choice among alternative conditional frames of reference when they exist, or by some radical interpretation of the consequences of (C), in which the role of experimental frames of reference in general in evidential interpretations is reappraised. The adoption of a convention to avoid certain possible applications of (C) would seem artificial and unsatisfactory in principle; on the other hand, the need for a radical reappraisal of the role of experimental frames of reference, which is apparent in the light of such examples, is confirmed quite generally by the above result, that (C) and (S) imply (L). For according to (L), reference to any particular experimental frame of reference, even an actual or a conditional one, for evidential interpretations, has necessarily a partly-conventional character.

Earlier proofs that (C) and (S) imply (L), restricted to relatively simple classes of experiments, utilized recognition of mixture structures in experiments [3], [4]. But in the above proof that (C) and (S) imply (L) for all classes of experiments, no existence of mixture structures in the experiments E, E', under consideration was required; the ancillary used there was constructed with the hypothetical mixture E^*. The conclusion (L) takes us beyond the need to examine specific experiments for possible mixture structure, since it eliminates the need to regard any experimental frames of reference, including actual or conditional ones, as essential for evidential interpretations. The possible usefulness of experimental frames of reference in a partly conventional sense for evidential interpretations will be discussed in some of the following sections.

Part II

6. *Evidential interpretations of likelihood functions.* We have seen above that on certain grounds, the likelihood principle (L) gives a solution of the first general problem of informative inference, that of mathematical characterization of evidential meaning. On this basis the second general problem of informative inference, that of evidential interpretations in general, can be described more precisely as the problem of evidential interpretations of likelihood functions. The remaining sections of this paper are devoted to the latter problem, that is, to consideration of questions like the following: When any instance (E, x) of statistical evidence is represented by just the corresponding likelihood function $L(\theta) = cf(x, \theta)$ (c an arbitrary positive constant), what are the qualitative and quantitative properties of the statistical evidence represented by $L(\theta)$? What concepts and terms are appropriate for describing and interpreting these evidential properties? How are such modes of evidential interpretation related to those in current general use?

The principal writers supporting the use of just the likelihood function for informative inference have not elaborated in very precise and systematic detail the nature of evidential interpretations of the likelihood function. Fisher has recently given a brief discussion and examples of such interpretations [9, especially pp. 68–73, 128–31]. He describes the relative likelihoods of alternative values of parameters as giving "a natural order of preference among the possibilities" (p. 38); and states that inspection of such relative likelihoods "shows clearly enough what values are implausible" (p. 71). Such interpretations were also recently discussed and illustrated by Barnard [2]. Both writers stress that point estimates, even when supplemented by measures of precision, have limited value for these purposes. For example when log $L(\theta)$ has (at least approximately) a parabolic form, then a point estimate (maximum likelihood) and a measure of its precision (preferably the curvature of log $L(\theta)$ at its maximum) constitute a convenient mode of description of the complete likelihood function (at least approximately); but more generally, with very different forms of $L(\theta)$, such descriptive indices have less descriptive value.

More detailed discussion of evidential interpretations of likelihood functions, and clarification of the meanings of terms appropriate for such discussion, seems desirable if possible, as has been remarked by Cox [8, p. 366]. These are the purposes of the following sections. Since any non-negative function $L(\theta)$, defined on an arbitrary parameter space, is a possible likelihood function, it is convenient to consider in turn parameter spaces of various forms, beginning for simplicity with the case of a two-point parameter space, followed by the case of any finite number of parameter points, and then more general and typical cases.

7. *Binary experiments.* (Parts of this section are closely related to reference [3, pp. 429–34].)

7.1. The simplest experiments, mathematically, are *binary* experiments, that is, experiments with parameter spaces containing just two points, θ_1, θ_2, representing just two simple hypotheses, H_1, H_2. Any outcome x of any such experiment determines a likelihood function $L(\theta)=cf(x, \theta)$ which may be represented by the pair of numbers $(cf(x, \theta_1), cf(x, \theta_2))$, with c any positive constant. Hence $L(\theta)$ is more parsimoniously represented by $\lambda=\lambda(x)=f(x, \theta_2)/f(x, \theta_1)$. ($\lambda(x)$ is the likelihood ratio statistic, which appears with rather different interpretations in other approaches to statistical theory.) Each possible likelihood function arising from any binary experiment is represented in this way by a number λ, $0\leq\lambda\leq\infty$. What sorts of evidential interpretations can be made of such a number λ which represents in this way an outcome of a binary experiment?

As a convenient interpretative step, consider for each number α, $0\leq\alpha\leq\frac{1}{2}$, a binary experiment whose sample space contains only two points, denoted "positive" (+) and "negative" (−), such that $\text{Prob}(+|H_1)=\text{Prob}(-|H_2)=\alpha$. Any such *symmetric simple* binary experiment is characterized by the "error probability" α which is the common value of "false positives" and "false negatives." (α is the common value of error-probabilities of Types I and II of the test of H_1 against H_2 which rejects just on the outcome +.) The outcomes of such an experiment determine the likelihood functions $\lambda(+)=(1-\alpha)/\alpha\geq1$ and $\lambda(-)=\alpha/(1-\alpha)=1/\lambda(+)\leq1$ respectively, with smaller error probabilities giving values farther above and below unity, respectively. According to the likelihood principle (L), when *any* binary experiment E gives *any* outcome x determining a likelihood function $\lambda(x)\geq1$, the evidential meaning of $\lambda(x)$ is the same as that of the positive outcome of the symmetric simple binary experiment with error-probability α such that $\lambda(x)=(1-\alpha)/\alpha$, that is, $\alpha=1/(1+\lambda(x))$. If the actual experiment E had the latter form, the outcome would customarily be described as "significant at the α level" (possibly with reference also to the Type II error-probability, which is again α). This currently standard usage can be modified in a way which is in accord with the likelihood principle by calling $\alpha=1/(1+\lambda(x))$ the *intrinsic significance level* associated with the outcome x, regardless of form of E. Here the probability α is defined in a specified symmetric simple binary experiment, and admits therein the usual conceptual frequency interpretations. The relations between such an experiment and the outcome $\lambda(x)$ being interpreted are conceptual, in a way which accords with the likelihood principle; the conventional element involved in adopting such an experimental frame of reference for evidential interpretations is clear, and is necessary in the light of the likelihood principle. (Alternative conventions of choice of experimental frames of reference are discussed in reference [3].) Outcomes giving $\lambda(x)\leq 1$ can be interpreted similarly: such outcomes support H_1 against H_2, with evidential strength corresponding to the intrinsic significance level $\alpha=\lambda(x)/(1+\lambda(x))$.

In connection with the current use of significance levels in evidential interpretations, it has often been stressed that consideration of the power of tests is essential to reasonable interpretations. But no systematic way of considering power along with significance levels seems to have been proposed specifically for the purpose of informative inference. And current standard practice often fails to include such consideration in any form (cf. reference [12]). The concept of intrinsic significance level incorporates automatic consideration of error-probabilities of both types, within its own experimental frame of reference, in a way which is also in accord with the likelihood principle.

7.2. Tukey [22] has recently stressed the need for a critical reappraisal of the role of significance tests in the light of a history of the practice and theory of informative inference. The next paragraphs are a brief contribution in this direction from the present standpoint.

Because the function of informative inference is so basic to empirical scientific work, it is not surprising that its beginnings can be traced back to an early stage in the development of the mathematical theory of probability. As early as 1710, Dr. John Arbuthnot computed the probability of an event which had been observed, that in each of a certain 82 successive years more male than female births would be registered in London, on the hypothesis that the probability of such an event in a single year was $\frac{1}{2}$; and he interpreted the very small probability $(\frac{1}{2})^{82}$ as strong evidence against the hypothesis [19, pp. 196–8]. This was perhaps the earliest use of a formal probability calculation for a purpose of statistical inference, which in this case was informative inference. Other early writers considered problems involving mathematically similar simple statistical hypotheses, and alternative hypotheses of a statistically-degenerate kind under which a particular outcome was certain: a "permanent cause" or "certain cause," or non-statistical "law of nature," that is, a hypothesis "which always produces the event" [6, pp. 261, 358]. (It is not altogether clear that a simple non-statistical alternative would correspond to

Arbuthnot's view of his problem.) Non-occurrence of such an outcome, even once in many trials, warrants rejection of such a hypothesis without qualification or resort to statistical considerations; but occurrence of such an outcome on each of n trials provides statistical evidence which requires interpretation as such. If the event in question has probability p of occurrence in one trial under the first hypothesis (and probability 1 under the second), then the probability of its occurrence in each of n independent trials is $P=p^n$ under the first hypothesis (and 1 under the second). (It is convenient to assume in our discussion that n was fixed; this may be inappropriate in some interpretations of these early examples.) In Arbuthnot's example, $P=(\frac{1}{2})^{82}$.

In such problems, the quantity on which evidential interpretations center is P, and small values of P are interpreted as strong evidence against the first hypothesis and for the second. What *general* concepts and basic terms are involved in these simple and "obviously sound" evidential interpretations? We can distinguish three mathematical concepts which do not coincide in general, but which assume the common form P in cases of the present extreme simplicity: Here P is not only the probability of "what was observed" under H_1: (a) P is the probability of an outcome "at least as extreme as that observed" under H_1 (because here there are no outcomes which are "more extreme"); that is P is a *significance level* (or critical level); and (b) P is the ratio of the probabilities, under respective hypotheses, of "what was observed"; that is, P is a *likelihood ratio* λ. To determine whether (a) or (b) is the *appropriate* general concept of evidential interpretation which is represented here by the obviously-appropriate quantity P, we must turn to more general considerations, such as the analysis of the preceding sections. Since in more complex problems the two concepts no longer coincide, one may wonder whether early and current uses of the significance level concept have sometimes derived support by inappropriate generalization, to (a) as against (b), from such simple and perhaps deceptively "clear" examples.

7.3. It is convenient to discuss here a reservation sometimes expressed concerning (L) itself, because this reservation involves significance levels. Experiments of different structures, for example experiments based on observations of the same kind but based on different sampling rules, may lead to respective outcomes which determine the same likelihood function but which are assigned different significance levels according to common practice. It is felt by many that such differences in significance levels reflect genuine differences between evidential meanings, corresponding to the different sampling rules; and therefore that (L) is unreasonable because it denies such differences of evidential meaning. The following discussion of a concrete example may throw further light on this point, while providing additional illustrations of (C) and (L) and their significance. Consider once more the binomial experiment E_B of Section 4.4 above, consisting of three independent observations on Y, which takes the values 0 or 1, with probabilities .9, .1, respectively under H_1, and with probabilities .1, .9, respectively under H_2. Consider also a sequential experiment E_S in which independent observations of the same kind Y are taken until for the first time $Y=0$ is observed: Let Z denote the number of times $Y=1$ is observed before termination of such an experiment. Then the distribution of Z is given by $f_1(z)=(.9)(.1)^z$, under H_1, and by $f_2(z)=(.1)(.9)^z$, under H_2, for $z=0$, 1, $2 \cdots$. The experiment E_S can be represented as a mixture of simple binary component experiments, among which is the component E_2 (described in Section 4.3) consisting of a single observation Y; this component is assigned probability .09 in the mixture experiment equivalent to E_S. We recall that E_B also admits a mixture representation, in which the component E_2 appears, assigned probability .27. We may imagine two experimenters, using E_B and E_S respectively for the same purpose of informative inference, and we may imagine a situation in which the mathematical component experiments are realized physically by alternative measuring instruments as in our discussion of E_B in Section 4.3. Then the first experimenter's design E_B includes the equivalent of a .27 chance of using the instrument represented by E_2 (for a single observation); and the second experimenter's sequential design E_S includes the equivalent of a .09 chance of using the same instrument (for one observation). If by chance each experimenter obtained this instrument and observed a positive outcome from it, then evidently the two results would have identical evidential meaning (as (C) asserts). However the customary assignment of significance levels would give such results the .028 significance level in the framework of E_B, and the .01 significance level in the framework of E_S. Both of these differ from the .10 error-probability which characterizes the common component experiment E_2. The latter value would be the intrinsic significance level assigned in the interpretation suggested above; this value would be indicated immediately, in any of the experimental frames of reference mentioned, by the common value 9 assumed by the likelihood ratio statistic λ on each of the outcomes mentioned.

8. *Finite parameter spaces.* If E is any experiment with a parameter space containing only a finite number k of points, these may conveniently be labeled $\theta=i=1, 2, \cdots, k$. Any observed outcome x of E determines a likelihood function $L(i)=cf(x, i)$, $i=1, \cdots, k$. We shall consider evidential interpretations of such likelihood functions in the light of the likelihood principle, in cases where

$$\sum_{i=1}^{k} f(x, i)$$

is positive and finite. (The remaining cases are special and artificial in a sense related to technicalities in the role of density functions in defining continuous distributions.) It is convenient here to choose c as the reciprocal of the latter sum, so that without loss of generality we can assume that

$$\sum_{i=1}^{k} L(i) = 1.$$

The present discussion formally includes the binary case, $k=2$, discussed above.

Any experiment E with a finite sample space labeled $j=1, \cdots, m$, and finite parameter space is represented conveniently by a stochastic matrix

$$E = (p_{ij}) = \begin{bmatrix} p_{11} & \cdots & p_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ p_{k1} & \cdots & p_{km} \end{bmatrix},$$

where

$$\sum_{j=1}^{m} p_{ij} = 1,$$

and $p_{ij} = \text{Prob}[j|i]$, for each i, j. Here the ith row is the discrete probability distribution p_{ij} given by parameter value i, and the jth column is proportional to the likelihood function $L(i) = L(i|j) = cp_{ij}$, $i = 1, \cdots, k$, determined by outcome j. (The condition that

$$\sum_{i=1}^{k} p_{ij}$$

be positive and finite always holds here, since each p_{ij} is finite, and since any j for which all $p_{ij} = 0$ can be deleted from the sample space without effectively altering the model E.)

8.1. Qualitative evidential interpretations. The simplest nontrivial sample space for any experiment is one with only two points, $j = 1, 2$. Any likelihood function $L(i)$ (with

$$\sum_{i=1}^{k} L(i) = 1,$$

which we assume hereafter) can represent an outcome of such an experiment, for we can define

$$\text{Prob}[j = 1 \mid i] = L(i) \quad \text{and} \quad \text{Prob}[j = 2 \mid i] = 1 - L(i),$$

for $i = 1, \cdots, k$.

For example, the likelihood function $L(i) \equiv \frac{1}{3}$, $i = 1, 2, 3$, represents the possible outcome $j = 1$ of the experiment

$$E = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

Since this experiment gives the same distribution on the two-point sample space under each hypothesis, it is completely uninformative, as is any outcome of this experiment. According to the likelihood principle, we can therefore conclude that the given likelihood function has a simple evidential interpretation, regardless of the structure of the experiment from which it arises, namely, that it represents a completely uninformative outcome. (The same interpretation applies to a constant likelihood function on a parameter space of any form, as an essentially similar argument shows.)

Consider next the likelihood function $(1, 0, 0)$. (That is, $L(1) = 1$, $L(2) = L(3) = 0$, on the 3-point parameter space $i = 1, 2, 3$.) This represents the possible outcome $j = 1$ of the experiment

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

The outcome $j=1$ of E is impossible (has probability 0) under hypotheses $i=2$ and 3 (but is certain under $i=1$). Hence, its occurrence supports without risk of error the conclusion that $i=1$. According to the likelihood principle, the same certain conclusion is warranted when such a likelihood function is determined by an outcome of *any* experiment. (Similarly any likelihood function which is zero on a parameter space of any form, except at a single point, supports a conclusion of an essentially non-statistical, "deductive" kind.)

The likelihood function $(\frac{1}{2}, \frac{1}{2}, 0)$ could have been determined by outcome $j=1$ of

$$E = \begin{Bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{Bmatrix}.$$

This outcome of E is impossible under hypothesis $i=3$, and hence supports without risk of error the conclusion that $i\neq 3$ (that is, that $i=1$ or 2). Furthermore, E prescribes identical distributions under hypotheses $i=1$ and 2, and hence the experiment E, and each of its possible outcomes, is completely uninformative as between $i=1$ and 2. The likelihood principle supports the same evidential interpretations of this likelihood function regardless of the experiment from which it arose. (Parallel interpretations show that in the case of any parameter space, any bounded likelihood function assuming a common value on some set of parameter points is completely uninformative as between those points.)

In the preceding experiment, the distinct labels $i=1$ and 2 would ordinarily be used to distinguish two hypotheses with distinct physical meanings, that is, two hypotheses about some natural phenomenon which could be distinguished at least in a statistical sense by a suitably designed experiment. The particular experiment E is, as mentioned, completely uninformative as between these hypotheses. Therefore if an experiment of the form E were conducted, it would be natural for some purposes to describe the actual experimental situation in terms of a two-point parameter space, labeled by $i'=1$ or 2, and by the model

$$E' = (p'_{i'j}) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}.$$

Here $i'=2$ stands just for the same simple hypothesis previously denoted by $i=3$ in E; $i'=1$ represents a simple (one-point) hypothesis in this actual experimental situation, but also represents the composite hypothesis previously denoted by $i=1$ or 2 in E. Such examples illustrate a sense in which even the specification of the number of points in the parameter space (of an adequate mathematical-statistical model of an experiment) sometimes involves an element of conventionality.

Consider the likelihood function $(.8, .1, .1)$ on the 3-point parameter space $i=1, 2, 3$. The interpretation that this likelihood function (or the outcome it represents) has the qualitative evidential property of *supporting* the hypothesis $i=1$, against the alternatives $i=2$ or 3, is supported by various considerations including the following: This likelihood function represents the outcome $j=1$ of

$$E = \begin{pmatrix} .8 & .2 \\ .1 & .9 \\ .1 & .9 \end{pmatrix} = (p_{ij}).$$

With use of E, if one reports the outcome $j=1$ as "supporting $i=1$" (in a qualitative, merely statistical sense), and if one reports the remaining outcome differently, for example as "not supporting $i=1$," then one makes inappropriate reports only with probability .1 when $i=2$ or 3, and only with probability. 2 if $i=1$. (Without use of an informative experiment, such reports could be arrived at only arbitrarily, with possible use of an auxiliary randomization variable, and the respective probabilities of inappropriate reports would then total unity.) This illustrates, in the familiar terms of error-probabilities of two kinds defined in the framework of a given experiment, the appropriateness of this qualitative evidential interpretation. According to the likelihood principle, the same qualitative interpretation is appropriate when this likelihood function is obtained from any experiment. (It can be shown similarly that on any parameter space, when any bounded likelihood function takes different constant values on two respective "contours," each point of the contour with greater likelihood is supported evidentially more strongly than each point with smaller likelihood.)

Consider the respective likelihood functions (.8, .1, .1) and (.45, .275, .275); the latter is "flatter" than the first, but qualitatively similar. The interpretation that the first is *more informative than* the second (and therefore that the first supports $i=1$ more strongly than the second) is supported as follows: Consider

$$E = \begin{pmatrix} .8 & .2 \\ .1 & .9 \\ .1 & .9 \end{pmatrix} = (p_{ij}).$$

Consider also the experiment E' based on E as follows: When outcome $j=2$ of E is observed, an auxiliary randomization device is used to report "$w=1$" with probability $\frac{1}{2}$, and to report "$w=2$" with probability $\frac{1}{2}$; when outcome $j=1$ of E is observed, the report "$w=1$" is given. Simple calculations verify that E' has the form

$$E' = \begin{pmatrix} .9 & .1 \\ .55 & .45 \\ .55 & .45 \end{pmatrix} = (p'_{iw}).$$

The outcome $w=1$ of E' determines the likelihood function (.45, .275, .275) given above (the latter is proportional to the first column of E'). The experiment E' is less informative that E, since it was constructed from E by "adding pure noise" (randomizing to "dilute" the statistical value of reports of outcomes). In particular, the outcome $w=2$ of E' is exactly as informative as the outcome $j=2$ of E, since $w=2$ is known to be reported only when $j=2$ was observed. But the outcome $w=1$ of E' is less informative that the outcome $j=1$

of E, since $w=1$ follows all outcomes $j=1$ of E and some outcomes $j=2$ of E.

The preceding example illustrates that some likelihood functions on a given parameter space can be compared and ordered in a natural way. It can be shown that some pairs of likelihood functions are not comparable in this way, so that in general only a partial ordering of likelihood functions is possible. (An example is the pair of likelihood functions $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ and $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$.) The special binary case, $k=2$, is simpler in that all possible likelihood functions admit the simple ordering corresponding to increasing values of λ.

8.2. Intrinsic confidence methods. (Parts of the remainder of this paper where finite parameter spaces are considered are closely related to reference [4].) Consider the likelihood function (.90, .09, .01) defined on the parameter space $i=1, 2, 3$. This represents the possible outcome $j=1$ of the experiment

$$E = \begin{Bmatrix} .90 & .01 & .09 \\ .09 & .90 & .01 \\ .01 & .09 & .90 \end{Bmatrix} = (p_{ij}).$$

In this experiment, a confidence set estimator of the parameter i is given by taking, for each possible outcome j, the two values of i having greatest likelihoods $L(i|j)$. Thus outcome $j=1$ gives the confidence set $i=1$ or 2; $j=2$ gives $i=2$ or 3; and $j=3$ gives $i=3$ or 1. It is readily verified that under each value of i, the probability is .99 that the confidence set determined in this way will include i; that is, confidence sets determined in this way have confidence coefficient .99. For those who find confidence methods a clear and useful mode of evidential interpretation, and who also accept the likelihood principle, it may be useful for some interpretive purposes to consider the given likelihood function, regardless of the actual experiment from which it arose, in the framework of the very simple hypothetical experiment E in which it is equivalent to the outcome $j=1$, and where it determines the 99 per cent confidence set $i=1$ or 2. According to the likelihood principle, considering this outcome in the hypothetical framework E does not alter its evidential meaning; moreover, any mode of evidential interpretation which disallows such consideration is incompatible with the likelihood principle. Of course the standard meanings of confidence sets and their confidence levels are determined with reference to actual experimental frames of reference (or sometimes actual-conditional ones) and not hypothetically-considered ones. Hence in the present mode of evidential interpretation, the hypothetical, conventional role of the experimental frame of reference E must be made clear. This can be done by use of the terms "intrinsic confidence set" and "intrinsic confidence coefficient (or level)" to refer to confidence statements based in this way on a specified conventionally-used experimental frame of reference such as E.

With the same experiment E, if for each j we take the single most likely parameter point, namely $i=j$, we obtain a one-point confidence set estimator with intrinsic confidence coefficient .90. Thus the given likelihood function, arising from an experiment of any form, determines the intrinsic confidence set $i=1$, with intrinsic confidence coefficient .90; the latter terms, again, are fully defined only when the form of the conventionally-used experiment E is indicated.

The general form of such intrinsic confidence methods is easily described as follows, for any likelihood function $L(i)$ defined on a finite parameter space $i=1, \cdots, k$, and such that

$$\sum_{i=1}^{k} L(i) = 1:$$

If there is a unique least likely value i_1 of i (that is, if $L(i_1)<L(i)$ for $i\neq i'$), let $c_1=1-L(i_1)$. Then the remaining $(k-1)$ parameter points will be called an intrinsic confidence set with intrinsic confidence coefficient c_1; if there is no unique least likely value of i, no such set will be defined (for reasons related to the earlier discussion of points with equal likelihoods). If there is a pair of values of i, say i_1, i_2, with likelihoods strictly smaller than those of the remaining $(k-2)$ points, call the latter set of points an intrinsic confidence set, with intrinsic confidence level $c_2=1-L(i_1)-L(i_2)$. And so on. The experiment in which such confidence methods are actual as well as intrinsic confidence methods will always be understood to be

$$E = \begin{Vmatrix} L(1) & L(k) \cdots & L(2) \\ L(2) & L(1) & L(3) \\ L(3) & L(2) & \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ L(k) & L(k-1) & L(1) \end{Vmatrix}.$$

E is determined uniquely from the given $L(i)$ by taking the latter to determine the respective first-column elements, and then by completing E so that it is a "cyclic-symmetric" matrix, as illustrated (satisfying $p_{ij}=p_{i-1,j-1}$ for all i, j, with a subscript i or $j=0$ here replaced by the value k).

By using here the basic technical relations between (ordinary) confidence methods and significance tests, we can obtain certain interpretations of the hypothesis-testing form from intrinsic confidence methods. For example, if a simple hypothesis $i=1$ is of interest, and if a likelihood function $L(i)$ from any experiment leads to an intrinsic .99 confidence set containing $i=1$, the outcome can be interpreted as "not intrinsically significant at the .01 level." If the same likelihood function determines an intrinsic .95 confidence set not containing $i=1$, this can be interpreted as "intrinsically significant at the .05 level," or "supporting rejection of $i=1$ at the .05 intrinsic significant level." Here, in contrast with the special binary case $k=2$, a single interpretive phrase like the latter does not incorporate unambiguous reference to the power of a corresponding test defined in E; nor does a single intrinsic confidence set report automatically incorporate such reference. On the other hand, a report of the set of all intrinsic confidence sets, with their respective levels, as defined above, does incorporate such reference, for it is readily seen that such a report determines uniquely the form of the likelihood function which it interprets. (Systematic use of confidence methods rather than significance tests, when possible, and of sets of confidence sets at various levels has been recommended by a number of recent writers; cf. [22], [23], [24], [15], [5] and references therein.)

An important category of problems is that involving several real-valued parameters, in which suitable estimates or tests concerning one of the parameters are of interest, the remaining parameters being nuisance parameters. Many such problems can be considered in miniature in the case of a finite parameter space, for example by labeling the parameter points by (u, v), $u=1, \cdots, k'$, $v=1, \cdots, k''$, giving $k=k'k''$ points in all. Then intrinsic confidence sets for the parameter u can be defined, despite presence of the nuisance parameter v, by a generalization of the preceding discussion which includes a more general scheme for defining convenient relatively simple conventional experimental frames of reference.

9. *More general parameter spaces.* Suppose that the parameter space of interest is the real line, and that one wishes to make interpretations of the interval-estimate type based on the outcome of some experiment. Let $L(\theta)$ denote the likelihood function determined by an outcome of that experiment, and consider the hypothetical experiment E which gives outcomes y, $-\infty < y < \infty$, having for each θ the probability density function $g(y, \theta) = cL(\theta - y)$ and the corresponding cumulative distribution

$$G(y, \theta) = c \int_{-\infty}^{y} L(\theta - u) du.$$

We restrict our discussion to likelihood functions for which

$$\frac{1}{c} = \int_{-\infty}^{\infty} L(\theta) d\theta$$

exists and is positive and finite, so that the preceding definitions are applicable. In such an experiment E, θ is called a translation-parameter, since increasing θ shifts the density function to the right without altering its shape. In this experiment, the outcome $y=0$ determines the given likelihood function $L(\theta)$.

In any such experiment, a standard technique for construction of confidence limits is the following: For each y, let $\theta'(y)$ be a value of θ satisfying $G(y, \theta) = .95$, and let $\theta''(y)$ be a value of θ satisfying $G(y, \theta) = .05$. Then $\theta'(y)$ and $\theta''(y)$ are (ordinary) .95 confidence limits, and together they constitute a .90 confidence interval, for estimation of θ in E; when $y=0$ is observed, these become $\theta'(0)$ and $\theta''(0)$. By definition, we take the latter to be respective .95 intrinsic confidence limits for θ, and we take these together to constitute a .90 intrinsic confidence interval for θ; these terms are taken to incorporate automatic reference to the translation-parameter experiment E defined as above on the basis of the given likelihood function $L(\theta)$. Certain generalizations from this case are obvious.

As in the case of finite parameter spaces, intrinsic confidence concepts include systematic unambiguous use of conveniently-chosen conventional experimental frames of reference. In the light of the likelihood principle, evidential meanings of likelihood functions cannot be altered when they are interpreted in hypothetical experimental frames of reference, and this provides a useful way of simplifying inessential features, and of relating unfamiliar problems of evidential interpretation to familiar ones where familiar methods can be applied. But it would be unfortunate if such methods as intrinsic confidence methods were adopted without sufficiently deep appreciation of the role played by the conventionally-chosen frames of reference. For example, under a simple nonlinear transformation of the parameter space such as $\theta^* = \theta^3$, the physical and mathematical meaning of inference methods is unchanged; but if the parameter space of points θ is replaced in the above discussion by points $\theta^* = \theta^3$, the likelihood function $L(\theta)$ must be replaced by the (evidentially equivalent) one $L^*(\theta^*) \equiv L^*(\theta^3) \equiv L((\theta^*)^{1/3})$. The latter function of θ^* has a different form from the function $L(\theta)$ of θ, and would lead in the above discussion to more or less different intrinsic confidence statements. In this connection one may therefore give considerations to scalings or labelings of the parameter space which are of particular interest in connection with the subject-matter of an experiment.

Another consideration, which will not always coincide closely with the first, is that the adoption of a suitable scaling of the parameter space may allow technical or formal simplicity in the application and interpretation of intrinsic confidence methods. For example, if $L(\theta)$ has the form of a normal density function (to within a constant c), then intrinsic confidence methods coincide formally with readily-applicable standard confidence methods for estimation of the mean of a normal distribution. Here the density function $g(y, \theta) = cL(\theta - y)$ is formally a normal density function of y, for each θ, with standard deviation which we shall denote by $\sigma_{\hat{\theta}}$. When $y = 0$ we have the actual likelihood function under consideration, $cL(\theta)$; let $\hat{\theta}$ denote the value of θ which maximizes this function. This is of course the maximum likelihood estimate, $\hat{\theta} = \hat{\theta}(x)$, where x is the outcome of any actual experiment which determines such a likelihood function. The translation-parameter experiment E defined as above now represents one in which a single observation y is taken from a normal distribution with known standard deviation $\sigma_{\hat{\theta}}$ and with unknown mean $E(Y) = \mu = \theta - \hat{\theta}(x)$. (Recall that our parameter θ is not in general the mean μ of Y in this derived experiment; in general these are different, by the amount $\hat{\theta}(x)$. Recall also that the latter is a known number determined from the given $L(\theta)$.) Finally, when $y = 0$ is observed, the classical estimate of μ is $\hat{\mu} = y = 0$, and correspondingly the classical estimate of the linear function of μ, $\theta = \mu + \hat{\theta}(x)$, is $\hat{\theta} = \hat{\theta}(x)$. In the framework of E, the outcome represented by the given $L(\theta)$ is thus easily interpreted in various standard ways, including confidence intervals; all of these are in a sense summarized by stating simply that θ has the estimate $\hat{\theta}(x)$ with the known standard error $\sigma_{\hat{\theta}}$, and that the estimation problem has the familiar structure of the simplest standard problem of estimation of a normal mean. To take advantage of the familiarity and simplicity of experiments like E, whenever a likelihood function $cL(\theta)$ has the form of a normal density function, we shall call the standard deviation of that distribution, denoted $\sigma_{\hat{\theta}}$, the *intrinsic standard error* of an estimate; and shall use this term in conjunction with the maximum likelihood estimate $\hat{\theta}(x)$ because in the framework of E the latter plays the role of the classical best estimate. (A parallel discussion can be given for the case of likelihood functions $cL(\theta)$ which have approximately the form of a normal density. The corresponding evidential interpretations are approximate in ways which depend in detail on the nature of the "closeness" of $L(\theta)$ to the normal form.) An easy calculation of $\sigma_{\hat{\theta}}$ is based on the observation that

$$\frac{-1}{\sigma_{\hat{\theta}}^2} = \frac{\partial^2}{\partial\theta^2} \log L(\theta).$$

An important class of problems involving nuisance parameters are those in which sampling is performed non-sequentially or sequentially from a normal distribution with unknown mean μ and unknown standard deviation σ. Standard confidence methods are not easily and efficiently applicable for estimation of the mean in the case of sequential sampling, but in the non-sequential case the standard methods based on the t-distribution are applicable (and known to be efficient in various senses). Now the form of the likelihood function determined in any such sequential experiment coincides with that obtainable in a simple non-sequential experiment; and the form of such a likelihood function is always determined by the three numbers (statistics) N = sample size (number of observations, whether or not sequential sampling was used), $\hat{\mu}$ = sample mean (the simple mean of all observations), and s^2, the sample variance (computed from all observations by the standard formula which gives unbiased estimates of variance in the non-sequential case). It is convenient to call the confidence methods, based on these statistics in the way which is standard for the non-sequential case, *intrinsic confidence methods* for estimation of the mean of a normal distribution with unknown variance. Clearly any mode of evidential interpretation which would interpret such a set of statistics differently in the sequential and non-sequential cases is incompatible with the likelihood principle. The general approach illustrated here can clearly be readily applied in many other classes of problems. It will perhaps bear repetition, for emphasis, that all such methods as intrinsic significance levels, intrinsic confidence methods, and intrinsic standard errors, can, in the light of the likelihood principle, be nothing more than methods of expressing, in various ways, evidential meaning which is implicit in given likelihood functions.

10. *Bayesian methods: an interpretation of the principle of insufficient reason.* In the method of treating statistical inference problems which was initiated by Bayes and Laplace, it was postulated that some mathematical probability distribution defined on the parameter space, the "prior distribution," represents appropriately the background information, knowledge, or opinion, available at the outset of an experiment; and that this, combined with experimental results by use of Bayes' formula, determines the "posterior distribution" which appropriately represents the information finally available. This formulation is widely referred to as Bayes' principle (or postulate), and we shall denote it by (B). (In this general form it should perhaps be credited to Laplace.) The extra-mathematical content of this principle has been interpreted in several ways by various of its proponents as well as critics [11, pp. 6–12]. This approach in

general is not directed to the problem of informative inference, but rather to the problem of using experimental results along with other available information to determine an appropriate final synthesis of available information. However it is interesting to note that within this formulation the contribution of experimental results to the determination of posterior probabilities is always characterized just by the likelihood function and is otherwise independent of the structure of an experiment; in this sense we may say that (B) implies (L).

10.1. The principle of insufficient reason, which we shall denote by (P.I.R.), is the special case of (B) in which a "uniform prior distribution" is specified to represent absence of background information or specific prior opinion. Evidently the intention of some who have developed and used methods based on (P.I.R.) has been to treat, in suitably objective and meaningful terms, the problem of informative inference as it is encountered in empirical research situations. This case of (B) was of particular interest to early writers on Bayesian methods. Following Laplace, this approach was widely accepted during the nineteenth century. Analysis and criticism, notably by Boole [6] and Cournot, of the possible ambiguity of the notion of "uniformity" of prior probabilities, and of the unclear nature of "prior probabilities" in general, led later to a widespread rejection of such formulations. The principal contemporary advocate of this approach is Jeffreys [13].

It is at least a striking coincidence that when experiments have suitable symmetry (or analogous) properties, inference methods based upon (P.I.R.) coincide exactly in form (although they differ in interpretation) with various modern inference methods developed without use of prior probabilities. For example, if any experiment E with a finite parameter space happens to be cyclic-symmetric, then uniform prior probabilities ($1/k$ on each parameter point) determine posterior probability statements which coincide in form with ordinary confidence statements. As a more general example, if E' has a k-point parameter space but *any structure*, it is easily verified that such posterior probability statements coincide in form with the intrinsic confidence statements determined as in Section 8 above. It follows that, leaving aside questions of extra-mathematical interpretation of (P.I.R.) itself, (P.I.R.) can be taken as a formal algorithm for convenient calculation of intrinsic confidence statements in the many classes of problems where such agreement can be demonstrated.

When the parameter space is more general, the "uniform distribution" has usually been chosen as some measure which is mathematically natural, for example Lebesgue measure on a real-line parameter space, even when such a measure does not satisfy the probability axiom of unit measure for the whole (parameter) space. In such cases again the posterior probabilities determined by formal application of Bayes' formula agree in form with ordinary or conditional confidence statements when an experiment has suitable symmetry-like (translation-parameter) properties; and more generally, such posterior probability statements agree in form with the intrinsic confidence statements described in Section 9 above. Furthermore the questions of conventionality, concerning the specification of a "uniform" distribution in such cases, are exactly parallel in form to the features of conventionality of choice of experimental frame of reference discussed in Section 9.

10.2. A posterior probability statement determined by use of (P.I.R.) can be interpreted formally as merely a partial description of the likelihood function itself; and a sufficient number of such statements, or specification of the full posterior distribution, determine the likelihood function completely (provided the definition of "uniform" prior distribution is indicated unequivocally). This interpretation of (P.I.R.) makes it formally acceptable (in accord with (L)) as a solution of the first problem of informative inference, the mathematical characterization of evidential meaning. But this interpretation does not ascribe to (P.I.R.) any contribution to the second problem of informative inference, evidential interpretation, and does not include any specific interpretation of prior and posterior probabilities as such. On the interpretation mentioned, a posterior probability distribution might as well be replaced by a report of just the likelihood function itself. (On the basis of (L), without adoption of (P.I.R.) or (B), the absence of prior information or opinion admits a natural formal representation by a likelihood function taking any finite positive constant value, for example $L(\theta)\equiv 1$. Such a likelihood function is determined formally, for example, by any outcome of a completely uninformative experiment. Since likelihood functions determined from independent experiments are combined by simple multiplication, such a "prior likelihood function" combines formally with one from an actual experiment, to give the latter again as a final over-all "posterior" one.)

10.3. A more complete explication of (P.I.R.) is suggested by the close formal relations indicated above between intrinsic confidence statements and statements based on (P.I.R.). Writers who have recommended (P.I.R.) methods use the term "probability" in a broad sense, which includes both the sense of probabilities $\text{Prob}(A\,|\,\theta)$ defined within the mathematical model E of an experiment (which admit familiar conceptual frequency interpretations), and the sense in which any proposition which is supported by strong evidence is called "highly probable" (the latter sense, according to some writers, need not necessarily be given any frequency interpretation). It is in the latter sense that a high posterior probability seems to be interpreted by some writers who recommend (P.I.R.). Now the present analysis has led to the likelihood function as the mathematical characterization of statistical evidence, and to intrinsic confidence statements as a possible mode of evidential interpretation. In the latter, an intrinsic confidence coefficient plays the role of an index of strength of evidence; such a coefficient is determined in relation to probabilities defined in a mathematical model of an experiment (generally a hypothetical one), but such an index is not itself a probability *of* the confidence statement to which it is attached. However in the broad usage described above, such an index of strength of evidence can be called a probability. Such an index becomes *also* a (posterior) probability in the mathematical sense when a "uniform" prior distribution is specified; but we can alternatively regard the latter formalism as merely a convenient mathematical algorithm for calculating intrinsic confidence sets and their intrinsic confidence coefficients. Under the latter interpretation, the principle of insufficient reason does not constitute an extra-mathematical postulate, but stands just for a traditional mode of calculating and designating intrinsic confidence sets and their coefficients.

11. *An interpretation of Fisher's fiducial argument.* Fisher's program of developing a theory of fiducial probability is evidently directed to the problem of informative inference. This approach agrees with the traditional one based on the principle of insufficient reason, that statements of informative inference should have the form of probability statements about parameter values; but disagrees concerning appropriateness of adopting the principle of insufficient reason for determination of such statements (Fisher [9]). Such probabilities are defined by a "fiducial argument" whose full scope and essential mathematical structure have not yet been fully formalized. Nevertheless some of the mathematical and extra-mathematical features of this approach seem clear enough for discussion in comparison with the approaches described above.

In experiments with suitable symmetry (or analogous) properties, it has been recognized that fiducial methods coincide in form (although they differ in interpretation) with ordinary or conditional confidence methods. In more complex experiments such a correspondence does not hold; and Fisher has stated that in general fiducial probabilities need not be defined in an actual or actual-conditional experimental frame of reference, but in general may be defined in different conceptually-constructed but appropriate frames of reference. This fact, and the fact that symmetry (or mathematical transformation-group) properties of experimental frameworks play a prominent part in the fiducial argument, suggest that the frames of reference in which fiducial probabilities are to be considered defined may coincide in general with those in which intrinsic confidence methods are defined as in Sections 8 and 9 above.

The claim that fiducial probabilities are probabilities of the same kind discussed by the early writers on probability can perhaps be understood in the same general sense that "posterior probabilities" calculated under the principle of insufficient reason were interpreted in Section 10, that is, a high fiducial probability for a parameter set may be interpreted as an index of strong evidential support for that set. And the claim that such probabilities can be defined and interpreted independently of any extra-mathematical postulate such as (P.I.R.) could be interpreted in the same general sense as in the explication of (P.I.R.) suggested above in which the latter principle does not constitute an extra-mathematical postulate. In the latter interpretation, the fiducial argument would appear to be another purely mathematical algorithm for calculating statements of evidential interpretation.

These interpretations suggest that fiducial probability methods may in general coincide in form as well as in general intention with intrinsic confidence methods (and hence also with those based on (P.I.R.) as interpreted above); and that these approaches may differ only in their verbal and mathematical modes of expression.

The fiducial argument has usually been formulated in a way which does not apply to experiments with discrete sample spaces, nor to experiments lacking suitable symmetry properties. However, it is possible to formulate a version of the fiducial argument compatible with (L) which is free of such restrictions: If $E=(p_{ij})$ is any cyclic-symmetric experiment with a k-point parameter space, consider for each i the sufficient statistic

$$t(j,i) = \begin{cases} j-i+1, & \text{if the latter is positive,} \\ j-i+1+k, & \text{otherwise.} \end{cases}$$

When i is true, the corresponding statistic $t(j, i)$ has the distribution $\text{Prob}(t(J, i) = t \mid i) = p_{1t}$, $t=1, \cdots, k$. The form of the latter distribution is the same for each value of i, and hence can be written $\text{Prob}(t(J, i)=t)=p_{1t}$. (A family of statistics $t(j, i)$ with the latter property is a "pivotal quantity" in the usual terminology of the fiducial argument.) For each possible outcome j of E we define a mathematical probability distribution on the parameter space, the "fiducial distribution" determined by the observed value j, by

$$\text{Prob}(i \mid j) = \text{Prob}(t(j, I) = t) \equiv p_{1t}, \quad \text{where} \quad t = t(j, i).$$

Using the definition of $t(j, i)$ and the cyclic symmetry of E, this simplifies to

$$\text{Prob}(i \mid j) = p_{ij}.$$

Thus the fiducial distribution coincides here with the posterior distribution determined from (P.I.R.) and also with the likelihood function itself. The fiducial probability statements here will thus agree in form with posterior probability statements based on (P.I.R.) and also with ordinary confidence statements.

Next, suppose that E' is any experiment with a k-point parameter space, and consider the problem of evidential interpretations of an outcome of E' which determines a likelihood function $L(i)$. Under (L), the evidential meaning of $L(i)$ is the same as if $L(i)$ were determined by an outcome of a simple cyclic-symmetric experiment; and in the latter case, the fiducial statements determined as above would be formally available. Thus it seems appropriate to the general intention of the fiducial approach, and in accord with (L), to define the fiducial distribution by

$$L(i) \Bigg/ \sum_{i'=1}^{k} L(i')$$

where $L(i)$ is the likelihood function determined by any outcome of any experiment E' with a k-point parameter space, without restriction on the form of E'. Under this interpretation, the intrinsic confidence statements described in Section 8, and the posterior probability statements described in Section 10, would also correspond formally with fiducial probability statements. Perhaps similar correspondences can be traced in other classes of problems where the fiducial argument takes somewhat different forms.

12. *Bayesian methods in general.* As was mentioned in Section 10, Bayesian methods in general entail adoption of (L) for the delimited purpose of characterizing experimental results as actually used in such methods. In particular, for communication of any instance (E, x) of statistical evidence to one who will use or interpret it by Bayesian methods, it is sufficient (and in general necessary) to communicate just the corresponding likelihood function.

Much discussion of the differences between Bayesian methods in general and non-Bayesian statistical methods has centered on the likelihood principle. Hence it is of interest to consider here those distinctions and issues which may separate Bayesian methods in general (apart from (P.I.R.)) from methods and interpretations based on (L) but not (B). Such differences are not related to problems of informative inference, but concern problems of interpretation and/or use of likelihood functions, along with appropriate consideration of other aspects of an experimental situation including background ("prior") information, for scientific and/or utilitarian purposes.

Consider any binary experiment E concerning the statistical hypotheses H_1, H_2, in any situation of inference or decision-making where a certain "conclusion" or decision d would be adopted if the experimental outcome provides evidence supporting H_2 with sufficient strength. Apart from the simplicity of the binary case, evidently many inference situations can be described appropriately in such terms. Then it follows, from (L) and from the discussion of the evidential properties of the statistic λ in the binary case, that there is some critical value λ' such that the decision d would be adopted if and only if the outcome λ of E satisfies $\lambda \geq \lambda'$. The latter formulation can be recognized as appropriate and adopted, with some choice of λ' which seems appropriate in the light of the various aspects and purposes of the inference situation, along with some appreciation of the nature of statistical evidence as such; evidently this can be done by experimenters who adopt the likelihood principle but do not adopt Bayes' principle.

Consider alternatively, in the same situation, another experimenter whose information, judgments, and purposes are generally the same but who adopts and applies Bayes' principle. He will formulate his judgments concerning prior information by specifying numerical prior probabilities p_1, p_2, for the respective hypotheses H_1, H_2. He might formulate his immediate experimental purpose, if it is of a general scientific sort, by specifying that he will adopt the working conclusion d provided the posterior probability q_2 of d is at least as large as a specified number q_2'. Or if his experimental purpose is of a more utilitarian sort, he might specify that he will adopt the decision d provided that $q_2U_2 \geq q_1U_1$, where q_1, q_2 are respective posterior probabilities and U_1, U_2 are numerical "utilities" ascribed respectively to non-adoption of d when H_1 is true and to adoption of d when H_2 is true. Each such formulation leads mathematically to a certain critical value λ'' of the statistic λ and to an inference or decision rule of the form: Adopt d provided E yields an outcome $\lambda \geq \lambda''$. Thus there is no difference between the "patterns of inference or decision-making behavior" of Bayesian statisticians and of non-Bayesian statisticians who follow the likelihood principle, at least in situations of relatively simple structure. And, at least for such simple problems, one might say that (L) implies (B) in the very broad and qualitative sense that *use* of statistical evidence as characterized by the likelihood function alone entails that inference- or decision-making behavior will be externally indistinguishable from (some case of) a Bayesian mode of inference.

Some writers have argued that the qualitative features of the Bayesian mode of inference seem plausible and appropriate, but that the specification of definite numerical prior probabilities and the interpretation of specific numerical posterior probabilities seem less clearly appropriate and useful. (This viewpoint has been presented interestingly, with some detailed examples, by Polya [16].) The present writer hopes to see more intensive discussion, with detailed illustration by concrete examples, of the specific contributions which qualitative-Bayesian and quantitative-Bayesian formulations may have to offer to those statisticians who adopt the likelihood principle and interpret likelihood functions directly, making informal judgments and syntheses of the various aspects of inference or decision-making situations.

13. *Design of experiments for informative inference.* If an experiment is to be conducted primarily for purposes of informative inference, then according to (L) the various specific experimental designs E which are available are to be appraised and compared just in terms of the likelihood functions they will determine, with respective probabilities, under respective hypotheses, along with consideration of experimental costs of respective designs.

In the case of binary experiments, what is relevant is just the distribution of the statistic λ, defined in any binary experiment, under the respective hypotheses. The simplest specification of a problem of experimental design is evidently that a binary experiment should, with certainty, provide outcomes λ with evidential strength satisfying: $|\lambda| \geq \lambda'$, where λ' is a specified constant; for example, $\lambda' = 99$ indicates that each possible outcome of the experiment is required to have evidential strength associated (as in Section 7) with error-probabilities not exceeding .01. In experimental situations allowing sequential observation, it was shown in reference [3] that such a specification is met efficiently, in terms of required numbers of observations, by a design based on the sampling rule of Wald's sequential probability ratio test (with nominal error-probabilities both .01). If this sequential design is not feasible, some modification of the specified design criterion is indicated. For example, if only non-sequential designs are allowed, and a sample-size is to be determined, then in general one can guarantee only more or less high probabilities, under each hypothesis, that an experimental outcome will have at least the specified evidential strength.

Similarly, to obtain an intrinsic .95 confidence interval for the mean of a normal distribution with unknown variance, of length not exceeding a given number D, an efficient fully-sequential sampling rule is one which terminates when for the first time the .95 confidence interval, computed from all observations as if sampling were non-sequential, has length not exceeding D.

In general, such considerations concerning the design of experiments for informative inference under (L) lead to mathematical questions whose answers will often be found within the mathematical structures of the statistical theories of Fisher, Neyman and Pearson, and Wald, although these theories are typically used and interpreted differently, even for purposes of informative inference. For example, the distributions of the statistic λ in any binary experiment (which under (L) are basic for experimental design but irrelevant to evidential interpretation) are represented mathematically by the "α, β curve," which represents the binary experiment, and is the focus of attention in the

Neyman-Pearson and Wald treatments of binary experiments. More generally, the power functions of various tests admit interpretations relevant to experimental design under (*L*). And Fisher's asymptotic distribution theory of maximum likelihood estimates can be interpreted, as Fisher has indicated, as describing the asymptotic distributions, under respective hypotheses, of the likelihood function itself (at least in an interval around its maximum).

Clearly the problems of experimental design under (*L*) are manifold and complex, and their fruitful formulation and solution will probably depend on increased interest in and use of likelihood functions as such. Some of these problems of experimental design coincide in form with design problems as formulated by Bayesian statisticians [17]. Thus there is scope for interesting collaboration here between statisticians with somewhat different over-all viewpoints.

REFERENCES

[1] Bahadur, R. R., "Sufficiency and statistical decision functions," *Annals of Mathematical Statistics,* 25 (1954), 423–62.

[2] Barnard, G. A., discussion of C. R. Rao's paper, "Apparent anomalies and irregularities in maximum likelihood estimation," *Bulletin of the International Statistical Institute,* 38 (1961).

[3] Birnbaum, A., "On the foundations of statistical inference; binary experiments," *Annals of Mathematical Statistics,* 32 (1961), 414–35.

[4] Birnbaum, A., "Intrinsic confidence methods," *Bulletin of the International Statistical Institute,* Vol. 39 (to appear), Proceedings of the 33rd Session of the I.S.I., Paris, 1961.

[5] Birnbaum, A., "Confidence curves: an omnibus technique for estimation and testing statistical hypotheses," *Journal of the American Statistical Association,* 56 (1961), 246–9.

[6] Boole, G., *Studies in Logic and Probability.* La Salle, Illinois: Open Court Publishing Company, 1952.

[7] Cohen, M. R. and Nagel, E., *An Introduction to Logic and Scientific Method.* New York: Harcourt, Brace and Company, 1934.

[8] Cox, D. R., "Some problems connected with statistical inference," *Annals of Mathematical Statistics,* 29 (1958), 357–72.

[9] Fisher, R. A., *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd, 1956.

[10] Fisher, R. A., "The comparison of samples with possibly unequal variances," *Annals of Eugenics,* 9 (1939), 174–80.

[11] Good, I. J., *Probability and the Weighing of Evidence.* New York: Hafner Publishing Company, 1950.

[12] Harrington, G. M., "Statistics' Logic," *Contemporary Psychology,* Vol. 6, No. 9 (September 1961), 304–5.

[13] Jeffreys, H., *Theory of Probability,* Second Edition. London: Oxford University Press, 1948.

[14] Lehman, E., *Testing Statistical Hypotheses.* New York: John Wiley and Sons, Inc., 1959.

[15] Natrella, M. G., "The relation between confidence intervals and tests of significance," *The American Statistician,* 14 (1960), No. 1, 20–22 and 38.

[16] Polya, G., *Mathematics and Plausible Reasoning,* Volume Two. Princeton: Princeton University Press, 1954.

[17] Raiffa, H. and Schlaifer, R., *Applied Statistical Decision Theory.* Boston: Division of Research, Harvard Business School, 1961.

[18] Savage, L. J., "The foundations of statistics reconsidered," *Proceedings of the Fourth*

Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1961.

[19] Todhunter, I., *A History of the Mathematical Theory of Probability.* New York: Chelsea Publishing Company, 1949.

[20] Tukey, J., "A survey of sampling from a contaminated distribution," *Contributions to Probability and Statistics,* Ed. by I. Olkin, *et al.* Stanford: Stanford University Press, 1960, 448–85.

[21] Tukey, J., "The future of data analysis," *Annals of Mathematical Statistics,* 33 (1962), 1–67.

[22] Tukey, J., "Conclusions vs. decisions," *Technometrics,* 2 (1960), 423–33.

[23] Wallace, D. L., "Conditional confidence level properties," *Annals of Mathematical Statistics,* 30 (1959), 864–76.

[24] Wallace, D. L., "Intersection region confidence procedures with an application to the location of the maximum in quadratic regression," *Annals of Mathematical Statistics,* 29 (1958), 455–75.

[25] Wilson, E. B., *An Introduction to Scientific Research.* New York: McGraw-Hill Book Company, 1952.