ALLAN BIRNBAUM*

THE NEYMAN–PEARSON THEORY AS DECISION THEORY, AND AS INFERENCE THEORY; WITH A CRITICISM OF THE LINDLEY–SAVAGE ARGUMENT FOR BAYESIAN THEORY

1. INTRODUCTION AND SUMMARY

The concept of a *decision*, which is basic in the theories of Neyman–Pearson, Wald, and Savage, has been judged obscure or inappropriate when applied to interpretations of data in scientific research, by Fisher, Cox, Tukey, and other writers. This point is basic for most statistical practice, which is based on applications of methods derived in the Neyman–Pearson theory or analogous applications of such methods as least squares and maximum likelihood. Two contrasting interpretations of the decision concept are formulated: *behavioral*, applicable to 'decisions' in a concrete literal sense as in acceptance sampling; and *evidential*, applicable to 'decisions' such as 'reject H_1' in a research context, where the pattern and strength of statistical evidence concerning statistical hypotheses is of central interest. Typical standard practice is characterized as based on the *confidence concept* of statistical evidence, which is defined in terms of evidential interpretations of the 'decisions' of decision theory. These concepts are illustrated by simple formal examples with interpretations in genetic research, and are traced in the writings of Neyman, Pearson, and other writers. The Lindley–Savage argument for Bayesian theory is shown to have no direct cogency as a criticism of typical standard practice, since it is based on a behavioral, not an evidential, interpretation of decisions.

2. TWO INTERPRETATIONS OF 'DECISIONS'

Statistical decision problems are the subject of major theories of modern statistics, and have been developed with great precision and generality on the mathematical side. But in the view of many applied and theoretical

Synthese **36** (1977) 19–49.

statisticians, the scope and interpretation of decision theories has remained obscure or doubtful in connection with interpretations of data in typical scientific research situations.

The reason for concern here is that most statistical methods applied to research data have been given their most systematic mathematical justification within the Neyman–Pearson theory; and that theory in turn has been given its most systematic mathematical development within the (non-Bayesian) statistical decision theory initiated by Wald. In this development the alternative statistical hypotheses which may be 'rejected' or 'accepted' on the basis of a testing procedure are identified with the respective 'decisions' appearing in the formal model of a decision problem.

Similarly, each confidence interval which may be determined by an estimation procedure is identified with one of the 'decisions' of a model. This leads to questions about the scope and interpretation of the 'decision' concept which have been discussed by a number of writers: In what sense, if any, it is appropriate to regard the results of typical scientific data analysis based on standard statistical methods of testing and estimation as decisions?

We shall treat this question in a way which is self-contained, and more systematic in some respects than previous discussions. Our intention is to complement and clarify previous discussions in certain respects, without attempting to review or summarize them. The interested reader is urged to read or re-read such earlier discussions, particularly those of Tukey (1960), Cox (1958, p. 354), and others cited below.

The terms 'decide' and 'decision' were used heavily by Neyman and Pearson in the series of joint papers which initiated their theory, notably in the preliminary exploratory paper of 1928, and in the 1933 paper in which problems of testing statistical hypothesis were first formulated in a way which can be regarded as a case of statistical decision problems.

A frequently cited ('paradigm') type of application of statistical decision theories and of the Neyman–Pearson theory is that of industrial acceptance sampling (Neyman and Pearson, 1936, p. 204; Wald, 1950, pp. 2–3): A lamp manufacturer must decide whether or not to place a batch of lamps on the market, on the basis of tests on a sample from the batch.

The simplest models of decision problems are characterized fully, for our present purposes of discussion, by schemas of the following form:

Simple hypotheses:	H_1,	H_2
Possible decisions:	d_1,	d_2
Error probabilities:	$\alpha = \text{Prob}\,[d_1 \mid H_1]$,	$\beta = \text{Prob}\,[d_2 \mid H_2]$

A simple hypothesis is any probability distribution which may be defined over the range of possible outcomes (the sample space) of an experiment or observational procedure.

For example, the lamp manufacturer may be interested in the simple hypothesis H_1 that a batch of lamps contains exactly 4% defective lamps, and in the alternative simple hypothesis H_2 that the batch contains exactly 10% defectives, possibly because a batch is considered definitely good if it has 4% or fewer defectives, and is considered definitely bad if it has 10% or more defectives.

For a given batch, his possible decisions are:

d_1: withhold the batch from the market; and

d_2: place the batch on the market.

The performance of any decision function (that is any rule for using data on a sample of lamps from the batch to arrive at a decision d_1 or d_2) is characterized fully, under H_1 and H_2, by the respective error probabilities α and β defined in the schema. (An example of a decision function here is the rule: Place the batch on the market if and only if fewer than 3 defectives are found in a random sample of 25 lamps.)

Consider the interpretation of the decisions d_1 and d_2 which appear in the schema, in its application to the problem of the lamp manufacturer.

When the manufacturer places a batch of lamps on the market, he performs an *action*. If he does so after considering also one or more alternative possible actions, as in our example, then he has taken a *decision* in favor of that action.

Here the terms 'decision' and 'action' refer to the *behavior* of the manufacturer in a simple direct and literal way. We shall use the term *behavioral interpretation* of the decision concept to refer to any comparably simple, direct, and literal interpretation of a 'decision' appearing in a formal model of a decision problem.[1]

The behavioral interpretation must be criticized and rejected, in the view of many investigators and statisticians, when such a schema and model are applied in a typical context of scientific research in connection with standard methods of data analysis. Convenient examples may be drawn from genetic linkage studies, which have the general scientific goal of extending knowledge of the 'chromosome map' which largely characterizes a species or strain in classical Mendelian genetics.[2]

Consider an investigator who judges that his linkage studies provide very strong evidence that two genetic loci lie on the same chromosome (with the usual appreciation that future studies could conceivably reverse his judgement); and who reports his conclusion, together with a summary of his data and his interpretation of it, based in part on use of a test determined by applying the Neyman–Pearson theory (as in Morton, 1955, or Smith, 1953, pp. 180–3), in a research journal.

His conclusion favoring the scientific hypothesis of linkage corresponds in some way to a 'decision' d_1 in a schema like that above, where now H_1 is the statistical hypothesis characterizing no linkage. It is the nature of this correspondence which we wish to examine carefully.

3. STATISTICAL EVIDENCE, AND ITS INADEQUATE REPRESENTATION BY THE 'DECISIONS' OF DECISION THEORY

The problem of testing statistical hypotheses is often described (in the Neyman–Pearson papers and elsewhere) as a problem of deciding whether or not to 'reject a statistical hypothesis' such as H_1 (e.g. Neyman and Pearson, 1928, p. 1; 1933, p. 291). This suggests the interpretation given by most writers who formulate problems of testing as decision problems:

d_1: reject H_1

d_2: do not reject H_1.

But this interpretation leads immediately to the question: What is the interpretation of 'reject H_1' in, for example, the situation of the investigator of our example who concluded that linkage was present?

Even if the geneticist uses typical terminology such as 'reject H_1, the hypothesis of no linkage,' neither he nor his colleagues understand that

he is making a decision in any literal and unqualified sense which could be given a behavioral interpretation closely comparable with that of the lamp manufacturer's decision in the example above.

Rather, the decision-like term 'reject' expresses here an interpretation of the statistical evidence, as giving appreciable but limited support to one of the alternative statistical hypotheses. This evidential interpretation of the experimental results is in principle based on a *complete* schema of the kind indicated above, even when this is only implicit.

In this essential respect, the identification suggested above between 'reject H_1' and the single element d_1 of the schema is inadequate, and is misleading when taken out of the context of the schema. Such cases of statistical evidence are adequately represented by symbols like

d_1^*: (reject H_1 for H_2, α, β)

and

d_2^*: (reject H_2 for H_1, α, β),

each of which carries an indication of the complete schema which serves as the conceptual frame of reference for the interpretation of statistical evidence here.

The symbols d_1^* and d_2^* represent in prototype typical interpretations and reports of data treated by standard statistical methods in scientific research contexts.

We shall use the term *evidential interpretation* of the decision concept to refer to such applications of models of decision problems; and we shall use the term *confidence concept* of statistical evidence to refer to such interpretations of statistical evidence.

In the view of this writer and some others, although typical applications of standard statistical methods in research are of the kind we have illustrated, the central concepts guiding such applications and interpretations (for which we have introduced the terms in italics above) have *not* been defined within any precise systematic theory of statistical inference. Rather, these concepts exist and play their basic roles largely implicitly and unsystematically, in guiding applications and interpretations of standard methods, and in guiding the development of new statistical methods. We shall not offer any precise theoretical account of these concepts, nor even claim that such an account can be given. Our aims are limited to

illustrating the existence and wide scope of the confidence concept, and clarifying some of its features.

The confidence concept seems to be in part a primitive intuitive concept of statistical evidence associated with schemas of the above kind, which may be expressed in the following prototypic formulation:

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

Examples. The following are simple examples of the confidence concept of statistical evidence. They may be thought of in the context of the investigation of genetic linkage described above. The interpretations of statistical evidence are expressed in the first person because they illustrate in simple cases the writer's own practice and thinking concerning statistical evidence, based in part on some experience as an independent interpreter of genetic data and developer of some new methods in Mendelian theory and data analysis (Birnbaum, 1972), as well as on extensive observation and analysis of general statistical practice and thinking. In my view these examples, and their interpretations in following sections, are typical of widespread statistical thought and practice, with the qualification that they are given here with a degree and style of explicit expression which is unusual. The interested reader will of course make an independent judgment about this.

The first person form is somewhat analogous to the usage of Savage (1954) whose Bayesian decision theory is developed from the standpoint of a generic rational person 'you'. In a following section these examples will be referred to in the course of a critical discussion of some assumptions of Savage's and Wald's decision theories.

Symbols of the form d_1^* and d_2^* introduced above are used to present the examples.

(1) I interpret

(reject H_1 for H_2, 0.06, 0.08)

as strong statistical evidence for H_2 as against H_1. Similarly I interpret

(reject H_2 for H_1 0.06, 0.08)

as strong statistical evidence for H_1 as against H_2.

(2) I interpret

(reject H_1 for H_2, 0, 0.2)

as conclusive evidence for H_2 as against H_1. Here the zero value of the error probability of the first kind indicates that the observational results are incompatible with H_1.

(3) I interpret

(reject H_1 for H_2, 0.01, 0.2)

as very strong statistical evidence for H_2 as against H_1.

(4) I interpret

(reject H_2 for H_1, 0, 0.2)

as weak statistical evidence for H_1 as against H_2. Here the relatively large value 0.2 of the error probability of the second kind suggests relative skepticism concerning this evidence against H_2.

(5) I interpret

(reject H_1 for H_2, 0.5, 0.5)

as worthless statistical evidence. It is no more relevant to the statistical hypotheses considered than is the toss of a fair coin, since the error probabilities (0.5, 0.5) also represent a model of a toss of a fair coin, with one side labeled 'reject H_1' and the other 'reject H_2'. If such a case arose in practice, our comments would lead us to judge the experiment, or at least the test adopted, to be worthless.

The distinction between the two interpretations of 'decision' may be epitomized (as Bernard Norton has pointed out) by contrasting the ordinary usages:[3]

behavioral:

'decide *to*' act in a certain way,

and

evidential:

'decide *that*' a certain hypothesis is true or is
– supported by strong evidence.

Concerning the different (pragmatist) identification of 'decide that A is true or well supported' with 'decide to act as if A is true or well supported', it will be clear from discussion above and below that we reject any such simple identification, and regard conclusions and statistical evidence as having autonomous status and value.

The preceding considerations were emphasized clearly though less formally by Cox (1958, p. 354) as follows:

> 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.

Further analysis of the distinctions between the two interpretations of the 'decisions' of decision theory is provided in those sections below which treat certain assumptions underlying Savage's and Wald's decision theories. In particular, it is shown that if one wishes to regard evidential *statements* represented, for example, by

d_1^*: (reject H_1 for H_2, 0.05, 0.05)

as '*decisions*' in a formal model of a decision problem, then certain basic assumptions of statistical decision theories are incompatible with certain basic properties and meanings of those evidential statements.

4. STATISTICAL EVIDENCE AS ONE AMONG SEVERAL CONSIDERATIONS REGARDING SUPPORT OF SCIENTIFIC CONCLUSIONS

As Tukey (1960) has emphasized, a conclusion reached in a scientific investigation, such as the conclusion of our geneticist that two loci are linked, requires not only

(a) statistical evidence of sufficient strength concerning the statistical hypotheses of interest.

In addition the investigator (or community of investigators) must judge

(b) the adequacy of the mathematical-statistical model, which serves as the conceptual frame of reference for the interpretation of the statistical evidence, to represent the research situation in relevant respects; and

(c) the compatibility with other knowledge and evidence of a conclusion that may be supported by statistical evidence provided by the current investigation (for example, strong statistical evidence against the statistical hypotheses representing no linkage).[4]

These important considerations prevent us from regarding a scientific conclusion as being determined in any simple or exclusive way by the statistical evidence which may support it.

The Neyman–Pearson theory introduced a kind of formal symmetry into the formulation of problems of testing statistical hypotheses, by requiring explicit specification of alternative statistical hypotheses and error probabilities of the second kind (e.g. H_2 and β in our schema) to complement the traditional specification (e.g. just H_1 and α in our schema).

But in many early and modern applications of statistical tests, there is a definite lack of symmetry in the status of the alternative statistical hypotheses considered, related to a lack of symmetry in the status or significance of corresponding scientific hypotheses or possible conclusions. For example in many cases one scientific hypothesis is regarded as established on the basis of current knowledge, or at least as acceptable or plausible, unless and until sufficiently clear and strong evidence against it appears. Clearly such considerations lie outside the scope of mathematical statistical models and statistical evidence in the sense discussed above, but rather in the scope of the scientific background knowledge and judgment referred to in (b) and (c) above.

In traditional formulations of testing problems which preceded the Neyman–Pearson theory and which continue to appear prominently in applied and theoretical statistics, in various applications it may be more or less plausible to suppose that there is implicit, though not explicit, reference to alternative statistical hypotheses and corresponding error probabilities, as an implicit part of the basis for choice and reasonable interpretation of a test statistic; and possibly to suppose also that there is

implicit if not explicit reference to possible alternative scientific hypotheses or possible conclusions corresponding to such implicit statistical hypotheses. The scope of the present paper does not extend to tests in such traditional formulations except to the extent that they may be regarded in an application as being interpreted at least in principle with plausible implicit, if not explicit, reference to some alternative statistical hypotheses. Such terms as 'standard statistical methods' and 'standard methods of testing statistical hypotheses', as used throughout this paper, must be understood with this important qualification to avoid confusion.

5. THE THEORETICAL AMBIGUITY OF THE NEYMAN–PEARSON THEORY

The Neyman–Pearson theory is interpretable in its *mathematical* form as a special restricted part of general statistical decision theory, as we have indicated above and will elaborate further below. As to the *extra-mathematical* interpretations and theory, which relate that mathematical form to applications, one may say that there are *two* Neyman–Pearson theories:

One is based on behavioral interpretations of the decision concept, and has been elaborated by Neyman in terms of his concept of inductive behavior as mentioned above. It is difficult or (in the view of the present writer and some others) impossible to discover or devise clear plausible examples of this interpretation in typical scientific research situations where standard methods are applied. (The interested reader will make an independent judgement about this, and may wish to consider the extensive and important contributions of Neyman himself to the interpretation of scientific data in several research areas.)

The second theory which makes use of the mathematical structure of the Neyman–Pearson theory is based on evidential interpretations of the 'decisions' in that theory, and has as its central concept what we have called the confidence concept of statistical evidence – a concept whose essential role is recognizable throughout typical research applications and interpretations of standard methods, but a concept which has not been elaborated in any systematic theory of statistical inference.

Since even the existence of this important distinction between two theoretical interpretations of the mathematical structure of the Neyman–Pearson theory is not very widely nor clearly appreciated, much of the obscurity and misunderstanding found in the statistical literature is not surprising. A simple step toward limiting this confusion and obscurity would be to make consistent use of terms which keep the distinction in view whenever necessary, such as 'confidence concept' and 'evidential' or 'behavioral' interpretation; and to avoid unqualified use, when ambiguity and confusion could result, of such standard terms as: the Neyman–Pearson theory (or approach, or school); and 'frequentist', 'objectivist', 'orthodox', 'classical', 'standard', and the like.

In the many applications where each interpretation seems to have some role, a sharp theoretical distinction between the two interpretations may have particular value in helping to clarify the purpose or purposes of the application and guide the adoption of appropriate methods. For example, new knowledge about a genetic linkage may have immediate value as a basis for the genetic counseling of a particular family. Here one can in principle consider two models of decision problems as having some scope in the same situation, one having 'decisions' interpreted in the literal behavioral sense (for example 'do not have another child' or 'do'); and the other model having 'decisions' with evidential interpretations (for example concerning statistical hypotheses related to possible scientific conclusions about genetic linkage).

Even if various details of the two models should correspond (for example the two decision functions adopted might, though they need not, be identical in form though different in kind of interpretation), the purposes and problems considered would be distinct, and hence properly characterized and treated by distinct theoretical concepts.

In other applications where there is a problem of decisions in the behavioral sense, one may seek conclusions (or strong statistical evidence) as a basis for making decisions judiciously. In such cases, if some formal model of a decision problem is considered to be an accurate model of the real situation in the relevant respects, one may argue that to consider conclusions (or statistical evidence) as such is at best superfluous, and at worst may distract from clear appreciation of the actual decision problem and accurate model. On the other hand, if it is not clear that any formal model of the decision problem has sufficient realism to be

applied, then development of new knowledge (conclusions or statistical evidence) may naturally be sought as a basis for making decisions.[5]

The second example of the 1936 paper of Neyman and Pearson involves explicit consideration of both conclusions and related decisions, but is discussed so briefly and incompletely that I am unable to interpret it from the standpoint of the preceding paragraphs. No other examples of applications were discussed in the joint papers. Thus the joint papers contain no discussion of an application in which a scientific conclusion was the sole or primary object of an investigation. Various writings of E. S. Pearson (notably 1937, 1947, 1962) discuss applications in which both conclusions and decisions (in the behavioral sense) are of interest, with conclusions sought as a basis for making decisions.

6. THE CONCEPTS OF TESTS AND DECISIONS IN THE 1933 PAPER OF NEYMAN AND PEARSON

The 1933 paper of Neyman and Pearson begins (pp. 141–2) with explicit concern about the meanings of concepts and methods of testing.

The authors discuss "What is the precise meaning of the words 'an efficient test of a hypothesis?' There may be several meanings."

No concept of an 'efficient test' had appeared in the preceding literature of testing, but the term 'efficient' had been introduced into mathematical statistics by Fisher in connection with his theory of estimation in the early 1920's.

Fisher's theory, with its striking mathematical power and conceptual depths and obscurities, stood in the background of the efforts of Neyman and Pearson to initiate a comparably systematic theory of tests, as they indicated in the introduction to their exploratory paper of 1928. Their plan to treat testing problems in an exact form (rather than by asymptotic approximations for the case of large samples, as Fisher had done) would eliminate some purely technical complications and thereby facilitate clarity concerning concepts such as 'efficient' or its analogues in a theory of tests.

On the side of applications, there was as much need for a systematic theory of tests as there had been for a more systematic theory of estimation, to guide investigators in choosing among alternative possible

tests in problems of increasing complexity, where the common sense which had guided traditional testing practice faltered. (Neyman and Pearson began their 1930 paper with discussion of Romanovsky's 1928 paper which had given new distribution theory for several statistics for a standard testing problem, pointing out the open basic problem of "determining which is the most appropriate one to use in any given case.")

The 1933 paper supplied a definition of 'an efficient test' which is clear on the mathematical side, and is neutral in relation to the contrasting behavioral and evidential interpretations of 'decision' discussed above.

An efficient test is defined as one in which the error probabilities (such as α and β in our schema) are minimized (jointly in some appropriate sense). Whether evidential or behavioral interpretations of 'decisions' are in view, such minimization of error probabilities would seem to be a clearly appropriate goal. No concept of an 'efficient test' has, even now, been proposed in terms of the earlier tradition of formulating testing problems (without reference to error probabilities under alternative hypotheses). In this sense one may say that it appears to have been 'necessary' to make some change in the traditional *mathematical* formulation of testing problems, as a basis for introducing a concept of an 'efficient test' which might guide applications and theoretical developments.

In any case, Neyman and Pearson met a problem of broad theoretical and practical scope by changing some of the terms of the problem, as original investigators have frequently done in all problem areas.[6]

Although some change in the mathematical formulation of testing problems seems to have been necessary, in the sense just indicated, the *theoretical* innovation of the Neyman–Pearson theory, the behavioral interpretation of tests, was not necessary in the following sense: An evidential interpretation has been associated with typical applications of tests in scientific research investigations in all periods of their use (which dates from 1710), without apparent discontinuity during the years following 1933 when the mathematical structure of the Neyman–Pearson theory became widely accepted as the new or improved mathematical basis for the theory of tests.

This observation suggests the questions: 'What roles or functions was the behavioral interpretation intended to serve?' and 'What functions has it served?' The joint papers suggest less than clear answers, while later

papers written by Neyman and Pearson separately suggest clearer answers which are different for the respective authors.

Although the 1933 paper begins, as we have noted, with concern about the meanings of concepts of testing, it discusses only a mathematical aspect of the meaning of an 'efficient test'; and the meaning of 'a test' (or a 'decision' such as 'reject H_1') is not discussed systematically with regard to extra-mathematical interpretations. Brief but clear and contrasting behavioral and evidential interpretations appear:

Behavioral: "Such a rule tells us nothing as to whether in a particular case H is true when" ... "accepted" ... "or false when" ... "rejected." ... "But ... if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have" analogous assurance concerning the frequency of rejections of H when it is false." (p. 142.)

Evidential: 1. In the "method of attack ... in common use ... If P were very small, this would generally be considered as an indication that the hypothesis, H, was probably false, and *vice versa*." (p. 141.)

2. "Let us now for a moment consider the form in which judgements are made in practical experience. We may accept or we may reject a hypothesis with varying degrees of confidence; or we may decide to remain in doubt. But whatever conclusion is reached the following position must be recognized. If we reject H_0, we may reject it when it is true; if we accept H_0, we may be accepting it when it is false, that is to say, when really some alternative H_t is true." (p. 146.)

The authors' attitude toward evidential interpretations is not made quite clear. The preceding quotation from p. 142 gives approvingly the behavioral interpretation of a test in the new mathematical formulation, as against the traditional "method of attack ... in common use" (traditional mathematical formulation, with evidential interpretation). But the quotation from p. 146 (in a discussion not linked by the authors with that of pp. 141–2) describes approvingly the evidential interpretation of a test in the new mathematical formulation.

An interpretation which would reconcile this apparent discrepancy is to regard the behavioral interpretation as not intended to apply in a situation of scientific research in any direct, literal, or concrete sense which would be incompatible with an evidential interpretation of the 'decisions' in question; but rather intended to apply in such a situation in a way which is heuristic or hypothetical, serving to explain the inevitably abstract theoretical meanings associated with the error probabilities,

formal 'decisions' such as 'reject H_1', and evidential interpretations based on a formal model of a decision problem (test). Thus *hypothetical* behavioral interpretations may be regarded as playing a role in the inner theoretical core of the confidence concept.[7]

This interpretation of the relation between behavioral and evidential interpretations seems close to that expressed by E. S. Pearson in various writings (in particular 1937, 1947, 1955, 1962). Professor Pearson has kindly permitted the following quotations from unpublished notes which he wrote in April 1974, as comments on an earlier draft of the present paper. (The terms 'behavioral' and 'evidential' do not appear in the original notes; in their places there appear the respective terms 'literal' and 'elliptical', which were used in the earlier version of the present paper.)

> [In the 1920's and 1930's] . . . my outlook as a practising statistician would have been what you term evidential. But to build such a structure one had to set out a mathematical theory which led to rules which, on the face of things, suggested a behavioral interpetation. . . . I think you will pick up here and there in my own papers signs of evidentiality, and you can say now that we or I should have stated clearly the difference between the behavioral and evidential interpretations. Certainly we have suffered since in the way the people have concentrated (to an absurd extent often) on behavioral interpretations . . .

It must happen frequently that a reader interested in an application where an evidential interpretation is appropriate, when he encounters a behavioral interpretation of a statistical method such as appears in many expository and theoretical works, supplies his own evidential *re*interpretation of the given behavioral interpretation if the writer has not supplied one, in order to relate the method cogently to his intended application and interpretation.

The 1920's and 1930's were a period of much critical concern with the meanings and possible meaningless of terms and concepts in the philosophy of science, psychology, and various other disciplines as well as in statistics. These concerns were usually pursued in terms of such doctrines as behaviorism, operationalism, or verificationism.

Various writers applied these criteria with varying degrees of stringency, greater stringency entailing smaller scope and importance for the theoretical and hypothetical concepts.

Perhaps the widest and most lasting influences of these doctrines have been heightened appreciation of both the values and the limitations of

such criteria for the analysis and development of a discipline, along with a balancing appreciation of the essential roles of theoretical, hypothetical, and perhaps even metaphysical concepts.

7. THE STATUS OF THE CONFIDENCE CONCEPT IN THEORY AND APPLICATIONS

As mentioned above, there is no precise mathematical and theoretical system which guides closely the wide use of the confidence concept in standard practice. (It is not clear that further developments can alter this situation. Cf. Birnbaum, 1969.) Rival theoretical approaches to the interpretation of research data (notably the likelihood and Bayesian approaches) offer attractive features of systematic precision and generality; but their basic concepts fail to satisfy those who prefer the confidence concept for the kind of theoretical objective control it provides over the error probabilities of interest (appearing in schemas like that above).[8]

The ad hoc aspects of the confidence concept are encountered in all applications, including that of testing genetic linkage discussed above. These aspects are related to its mathematical basis in the Neyman–Pearson theory as follows.

In a given problem of two simple hypotheses, the problem of minimization of error probabilities α and β (solved by Neyman and Pearson in 1933) leads not to a unique best test or decision function but to a family of best tests, each of which has the smallest possible value of β among all tests with the same (or smaller) value of α, including for example the following points (α, β) representing respective best tests:

$$(0.01, 0.05), (0.02, 0.02), \text{ and } (0.05, 0.01).$$

For a given application such as our linkage investigation, nothing in the confidence concept nor the Neyman–Pearson theory leads to a particular choice among these, yet choices of this kind are always made, implicitly if not explicitly, whenever the confidence concept is applied.

Another aspect of the ad hoc character of the confidence concept is its great potential flexibility in applications, which has not been very widely exploited. We may illustrate this in the preceding problem of two simple hypotheses, where three possible tests were considered. We may define a generalized kind of test of statistical hypotheses in terms of a formal

decision function taking three (rather than the usual two) possible values, as follows:
The decision function takes the possible values:

d_1: strong evidence for H_2 as against H_1

d_2: neutral or weak evidence

d_3: strong evidence for H_1 as against H_2.

It takes the value d_1 on those sample points where the test characterized by (0.01, 0.05) would reject H_1; it takes the value d_3 on those points where the test (0.05, 0.01) would accept H_1; and it takes the value d_2 on the remaining sample points. Such a 'three-decision' test requires a scheme of a new form to represent its more numerous error probabilities, which are defined as follows:

$$\alpha_1 = \text{Prob}\,(d_1|H_1)$$
$$= \text{probability of a major error of Type I}$$

$$\alpha_2 = \text{Prob}\,(d_2|H_1)$$
$$= \text{probability of a minor error of Type I}$$

$$\beta_1 = \text{Prob}\,(d_3|H_2)$$
$$= \text{probability of a major error of Type II}$$

$$\beta_2 = \text{Prob}\,(d_2|H_2)$$
$$= \text{probability of a minor error of Type II}$$

(It follows from the assumption that the original tests were best, that these error probabilities are minimized jointly in the usual sense. The ad hoc character of two-decision tests has not been eliminated, but reappears in such three-decision tests; and is illustrated once more by considering the possible alternative four-decision test which could be determined similarly by using also the test characterized by (0.02, 0.02) above.

In contrast the likelihood approach, and the technically related Bayesian approaches, are formally elegant, allowing intuitively plausible direct interpretations of all possible numerical values of the likelihood ratio statistic as indicating strength of statistical evidence in this problem.)

As other examples of methods for implementation of the confidence concept, outside the familiar categories of testing and of estimation by

confidence regions, we may mention nested confidence regions and related tests (e.g. Birnbaum, 1961; Dempster and Schatzoff, 1965; Stone, 1969); and methods for 'generalized testing' among three or more alternative statistical hypotheses and for classification (e.g. Birnbaum and Maxwell, 1960).

Among theoretical contributions specifically concerned with apparent difficulties or impossibilities in the way of giving a precise general theoretical treatment of the confidence concept and associated concepts, we may mention Barndorff–Nielsen (1971, 1973), Buehler (1959), Buehler and Fedderson (1963), Birnbaum (1969, 1970, 1972b), Cox (1971), and Durbin (1970).

The confidence concept depends in principle upon an extra-mathematical interpretation of the error probabilities which appear in schemas like that above, and this interpretation is usually described as a 'frequentist' or 'objectivist' interpretation; and the same terms are often used to describe the whole approach based on the confidence concept. The two theoretical interpretations of the 'decision' concept discussed above have analogues in interpretations of probabilities.

The term *propensity interpretation* has become widely used among philosophers in recent years to denote the kinds of 'objective' interpretation which seem appropriate and accurate for many theoretical terms in science, including probability. (See for example Mellor, 1971; Hacking, 1965; Braithwaite, 1954.) The confidence concept seems to call for this kind of interpretation of error probabilities, rather than any more directly (literal, operationalist, behavioristic) frequency interpretation, as we have indicated in earlier discussion of the confidence concept. On this view, criticisms of frequency interpretations of probability, as against propensity interpretations, are not relevant to the confidence concept.

(Presumably any rounded interpretation of probability in a scientific discipline would specify a role for concepts of statistical evidence, and perhaps also for the notion of 'practical certainty' associated with some applications, among the aspects of meaning associated with probability and related theoretical terms, such as 'genetic factor' in Mendelian genetics.)

We shall not attempt to survey the current status of the confidence concept in theory and applications. This would be a formidable task, since it would call for an account of the largely implicit interpretations of

standard statistical methods in a great variety of scientific research disciplines, and in a large and growing statistical literature including theoretical and expository works. It is hoped that the present paper will prove helpful to the interested reader as he makes his own observations and judgements concerning the nature of standard theoretical and applied statistical work in various disciplines.

8. OBJECTIONS TO A BASIC ASSUMPTION OF THE LINDLEY-SAVAGE ARGUMENT FOR BAYESIAN THEORY

One of the important and influential theoretical arguments for Bayesian theory is the Lindley–Savage argument. We shall show here that this argument has no direct relevance nor persuasive force, as an argument for Bayesian methods as against typical standard statistical practice with scientific research data, by showing that an assumption of the argument holds only for 'decisions' under behavioral interpretations, but not under the evidential interpretations which constitute standard statistical practice.

The argument is elementary, being formulated in terms of simple examples of tests (decision functions) like those above. The original somewhat informal accounts of the argument by Savage (1962, pp. 173–5) and Lindley (1971, p. 13–14) should be read by the interested reader. They are complemented by a formalized version of the argument, with additional discussion, in an appendix below.

The Lindley–Savage argument concerns judgements of preference or else indifference (equivalence) between alternative decision functions (tests) in problems of two simple hypotheses, with each decision function represented by a point $P = (\alpha, \beta)$ in the unit square, determined by its error probabilities α and β.

Our discussion will be based on some simple examples of statistical evidence given above, which we continue to express in the first person usage.

Examples. In some research situations I would strongly prefer to use a decision function (test) characterized by (0.05, 0.05) rather than one characterized by (0.1, 0).

In such situations I particularly value the guarantee, which is provided by use of (0.05, 0.05), that strong evidence will be obtained (either

supporting H_1 against H_2, or supporting H_2 against H_1). The use of (0.1, 0) allows the possibility that merely weak evidence, represented by (reject H_1 for H_2, 0.1, 0), will be obtained. For example, the knowledge in the background of a linkage investigation may include strong (though not conclusive) statistical evidence for the locations of all but one of the genetic factors which control a certain system of immune reactions; and the current investigation may have as its object just to determine whether the remaining factor lies on chromosome No. 1 or No. 2.

Let H_1 now stand for the hypothesis of linkage with another factor known to lie on No. 1, and H_2 the alternative hypothesis. In this situation I would avoid the risk of getting merely weak evidence by choosing (0.05, 0.05) rather than (0.1, 0); and would be able to complete the pattern of knowledge (chromosome map) of the system on a basis of consistently strong evidence.

Similarly, in some situations (including the same linkage investigation), I would prefer (0.05, 0.05) to (0, 0.1), for similar reasons.

In some situations (including the same linkage investigation) I would be indifferent as between (0.1, 0) and (0, 0.1), on grounds of their symmetry and of judgements of symmetry concerning the investigation in question.

This pattern of preferences may be summarized by

$$(0.05, 0.05) > (0.1, 0) \sim (0, 0.1),$$

where $>$ stands for 'is preferred to' and $\sim$ stands for 'is equivalent to.'

This pattern of preferences is incompatible with Assumption (II) of the Lindley–Savage argument as formulated in the appendix. (It is also incompatible with a basic premise underlying Wald's theory, as will be indicated in the next section.) This example suffices to illustrate that that assumption is not satisfied generally by the 'decision' concept associated with statistical tests as interpreted evidentially (not behaviorally) in typical research applications.

A different but analogous example incompatible with Assumption (II) is the preference pattern

$$(0.1, 0) \sim (0, 0.1) > (0.05, 0.05).$$

In some research situations I would have this preference pattern. In particular, if the knowledge in the background of a linkage investigation includes *conclusive* statistical evidence for the locations of all but one of

the factors which control certain immune reactions, then with certain scientific goals in view I would strongly prefer, rather than the guarantee of strong (but inconclusive) evidence provided by (0.05, 0.05), the uncertain possibility of completing with conclusive evidence the pattern of knowledge in question which is provided by either (0.1, 0) or (0, 0.1); and I would be indifferent as between them.

Assumption (II) expresses in one important way the concept of *rationality* (or *consistency*, or *coherence*) which is central to all statistical decision theories. Our criticism of this assumption and the concept it expresses may serve as a warning against oversimplified judgements of 'irrationality' (or 'inconsistency', or 'incoherence').

9. COMMENTS ON A BASIC PREMISE OF WALD'S DECISION THEORY

'Mixtures' of decision functions play important technical and theoretical roles in the development of Wald's (1950) statistical decision theory.

An example of a mixture is symbolized by

$$M = \tfrac{1}{2}(0, 0.1) + \tfrac{1}{2}(0.1, 0).$$

Here (0, 0.1) and (0.1, 0) represent as before two decision functions (tests), characterized by their respective pairs of error probabilities. The whole expression M stands for another decision function defined in terms of those two decision functions and an auxiliary randomization variable, say a toss of a fair coin, as follows: If the coin shows heads, the decision function (0, 0.1) is applied to the observed sample point; otherwise (0.1, 0) is applied.

To determine the error probabilities which characterize the decision function M, we find readily

$$(\alpha, \beta) = \tfrac{1}{2}(0, 0.1) + \tfrac{1}{2}(0.1, 0) = (0.05, 0.05).$$

(For example, under H_1 the respective error probabilities are 0, if (0, 0.1) is applied; and 0.1, if (0.1, 0) is applied; and each will be applied with probability $\frac{1}{2}$.)

The preceding discussion is based on a tacit assumption of a behavioral, and not a literal, interpretation of the decision functions considered.

One way of illustrating this is by reference to an example of the preceding section:

Suppose my preference pattern includes

$$(0, 0.1) \sim (0.1, 0) > (0.05, 0.05).$$

Then it is plausible that I may be indifferent also as between (0, 0.1) and *M*, since the latter will provide me with an application of (0, 0.1) or else an application of (0.1, 0) which I regard as equally satisfactory. But this implies that my preference pattern includes

$$M > (0.05, 0.05),$$

or, representing M now by its pair of error probabilities as determined above,

$$(0.05, 0.05) > (0.05, 0.05)$$

which is absurd.

The fallacy in the preceding discussion is that the preference pattern first assumed above arose in an example of evidential interpretations of 'decisions', while the calculation of the preceding paragraph was based on a behavioral interpretation. In particular, the preference for (0, 0.1) as against (0.05, 0.05) was based on a particularly high value ascribed to the possibility of statistical evidence symbolized by

$$(\text{reject } H_1 \text{ for } H_2, 0, 0.1),$$

in which the 'decision' ('reject H_1 for H_2') appears within the symbol for an evidential interpretation.

On the other hand, in the calculation of the error probability

$$\alpha = \tfrac{1}{2}(0) + \tfrac{1}{2}(0.1) = 0.05$$

above, we considered *just* the 'decision' ('reject H_1 for H_2'), *without* qualifications concerning the error probabilities which characterize the *different* respective decision functions (schemas) from which that 'decision' can result – that is, we tacitly interpreted that 'decision' behaviorally.

The general point illustrated is that while behavioral interpretations of 'decisions' may play a very valuable heuristic role in the mathematical development of the Neyman–Pearson and Wald theories, statistical

methods developed within those theories can and must be interpreted (or reinterpreted) with care when considered for possible use with evidential interpretations.

APPENDIX. ON THE LINDLEY–SAVAGE ARGUMENT FOR BAYESIAN THEORY[9]

The Lindley–Savage argument takes as its point of departure a recognized problem encountered whenever the (non-Bayesian) theories of Neyman–Pearson and Wald are to be applied, the problem illustrated above as one source of the *ad hoc* character of the confidence concept: that of choosing among the various best tests (decision functions) (α, β) available for a given application. The argument shows that if this problem of choice is treated 'rationally' (or 'consistently', or 'coherently') in a sense discussed above in Section 8, then 'you' are "a Bayesian, whether you thought you wanted to be or not . . . . Thus, the Bayesian position can be viewed as a natural completion, an overlooked step in the classical theory." (Savage, 1962, p. 175.)

The last comments refer to the final step of the argument, which may be illustrated in prototype as follows: Suppose you judge as equivalent, for a given application, three decision functions characterized respectively by

$$(0, 0.1), (0.05, 0.05), \text{ and } (0.1, 0).$$

Then . . . "you" are "a Bayesian, whether you thought you wanted to be or not . . . " in the sense that your preference behavior, in this context, is indistinguishable from that of a Bayesian; for example, a Bayesian who ascribes prior probabilities g_1 and g_2 respectively to H_1 and H_2, and losses L_1 and L_2 respectively to the errors of the first and second types, will also be indifferent as between those three decision functions, provided that $g_1L_1 = g_2L_2$. Such 'indistinguishability' represents an aspect of the behaviorist point of view which is basic to Savage's Bayesian decision theory. But clear and important distinctions of viewpoints are evident here from the standpoint of a non-Bayesian who may have a decision problem in the behavioral sense but who may wish to reach a conclusion (in the sense discussed above) as a basis for making a decision, perhaps because he regards no complete model of a decision problem, including loss functions, as clearly accurate. Important distinctions are clear also

from the standpoint of an investigator who has no decision problem except in the sense of evidential interpretations under the confidence concept, and finds no place in his thinking for loss functions nor Bayesian probabilities of statistical hypotheses, even if he may be indifferent in a given research context between three tests represented by the three points above.

The final step of the argument, just discussed in prototype, follows a more formalized argument whose conclusion is that 'you' have a preference pattern among tests (decision functions) characterized by indifference sets consisting of parallel line segments which cover the unit square of points (α, β) (and thus coinciding with certain Bayesian preference patterns), including for example PP' and QQ' in Figure 1. We discuss the

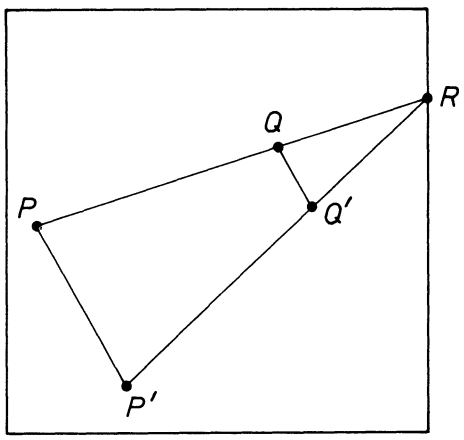

Fig. 1.

assumptions of this argument before presenting the derivation itself. The assumptions and derivation are formulated in terms of the mathematical concept of equivalence classes among points of the unit square; the interpretation of interest is that a person's indifference between two tests (decision functions) may be represented by stating that the points characterizing the tests are equivalent.

ASSUMPTION (I). There exist two distinct points P and P' which are equivalent.

Possible examples of (I) are P and P' in the figure; and the points $(0, 0.1)$ and $(0.1, 0)$ considered in examples above. This assumption seems free from possible plausible objections, for the following reason. The point $(0, 0)$ is preferred to the point $(0, 0.1)$, and the latter is preferred to $(0.1, 0.1)$, on the basis of the non-controversial principle of inadmissibility (regardless of possible evidential or behavioral interpretations which may be of interest). Consider the respective points (α, α) of the line segment from $(0, 0)$ to $(0.1, 0.1)$, and suppose that you judge that no such point is equivalent to $(0, 0.1)$.

Then as α increases continuously from 0, your preferences show an *implausible* discontinuity at some value of α, jumping from 'prefer (α, α) to $(0, 0.1)$' to 'prefer $(0, 0.1)$ to (α, α)' without anywhere assuming the intermediate value 'indifferent between (α, α) and $(0, 0.1)$'.

Our comments on the second assumption are stated conveniently with reference to a simpler restricted case:

ASSUMPTION (II*). If P and P' are equivalent, then P and P'' are also equivalent, where $P'' = kP + (1-k)P'$ and k is any number between 0 and 1.

For example if $k = \frac{1}{2}$, $P = (0, 0.1)$, and $P' = (0.1, 0)$, then $P'' = (0.05, 0.05)$, representing the example of a mixture discussed in Section 9 above, where we found the equivalence of $(0.05, 0.05)$ with $(0, 0.1)$ to be plausible under a behavioral interpretation of 'decisions' but not in general under an evidential interpretation. In particular we rejected that equivalence in the context of the examples of Section 8.

Assumption (II*) is the special case of the following assumption in which $R = P = Q$.

ASSUMPTION (II). If P and P' are equivalent, then Q and Q' are also equivalent, where $Q = kP + (1-k)R$, $Q' = kP' + (1-k)R$, R may be any point, and k may be any number in the unit interval.

LINDLEY–SAVAGE LEMMA. Assumptions (I) and (II) imply that the unit square is partitioned into equivalence sets, each consisting of a line segment parallel to PP'.

Proof:

(1) By (I) there exist two distinct equivalent points P and P'.

(2) Let R be any point on the perimeter of the unit square. Let k be any number satisfying $0<k<1$, and let $Q=kP+(1-k)R$, and let $Q'=kP'+(1-k)R$. (See Figure 1.) The case of R collinear with P and P' is mentioned below.) By (II), Q and Q' are equivalent, since P and P' are equivalent.

(3) The line segment QQ' is parallel with the segment PP', since the triangles RQQ' and RPP' are similar and have the common vertex R.

(4) Let c be any number satisfying $0<c<1$, and let $P''=cP+(1-c)P'$. P and P'' are equivalent, by (II).
(The special case (II*) of (II) applies here.)

(5) Since c is arbitrary, it follows that all points of the line segment PP' are equivalent. Similarly all points of the segment QQ' are equivalent.

(6) Since k is arbitrary, it follows that the triangle RPP' is covered by a family of line segments each parallel to PP', each of which is an equivalence class.

(7) As R sweeps out the circumference of the unit square, the square is covered by such triangles; and each triangle is again covered by segments parallel to PP', each segment consisting of equivalent points. (The case of R collinear with PP' is seen at this point not to be special.)

(8) The union of all such segments collinear with QQ' is a single segment between perimeter points of the square; since equivalence is transitive, this interval *is* an equivalence set. Similarly for other segments mentioned. Thus the unit square is partitioned into equivalence sets, each consisting of a line segment parallel to PP'.
This completes the proof of the Lemma.†

University College, London

† *Editors' Note.* The proofs of the present paper were ready only after the death of Professor Allan Birnbaum. The proofs were kindly checked by the staff of The City University, London and the University College, London. It was found that the bibliography was incomplete, and even though several corrections and additions were made, there still remain gaps in the bibliographical data.

NOTES

* The writer is grateful for helpful discussions of earlier versions of parts of this material with many colleagues, particularly D. V. Lindley, E. S. Pearson, J. Pratt, C. A. B. Smith, A. P. Dawid, G. Robinson, B. Norton, and M. Stone.

[1] The term 'rule of behavior' made its appearance, linked with the term 'decide', in the 1933 paper, in the discussion introducing the formulation of the problem of testing statistical hypotheses (p. 291, original; p. 142, reprint). Subsequently the concept of 'inductive behavior' was elaborated and supported, in opposition to various other concepts of statistical inference ('inductive reasoning'), by Neyman (1947, 1957, 1962, 1971).

[2] Among geneticists who are also prominent theoretical statisticians, the decision concept (at least in its behavioral interpretation) has been rejected as inappropriate in scientific data analysis, from different standpoints in statistical theory, by:

1. O. Kempthorne, from the standpoint of standard methods interpreted in a non-behavioral way similar to that discussed below (for example, 1971, pp. 471–3, 489);
2. C. A. B. Smith, who has developed a version of Bayesian theory, and has led in the use of Bayesian methods in scientific publications in genetics (1959, p. 297);
3. A. W. F. Edwards, an exponent of the likelihood approach, who has applied that approach in his scientific publications in genetics (1972); and
4. R. A. Fisher (for example, 1956, pp. 100–103).

The case of two simple hypotheses is unrealistic for problems of testing linkage, where a composite statistical hypothesis is generally adopted to represent the scientific hypothesis of linkage. However the simplified model of two simple hypotheses entails no sacrifice of realism with respect to the questions of interpretation considered in this paper. On the contrary, typical formulations of linkage tests in practice often make use of simple hypothesis, for technical reasons, to represent effectively a more realistic composite hypothesis (Morton, 1955; Smith, 1953, pp. 180–183).

Analogous comments apply to the limited realism of our discussion of the example of the lamp manufacturer: It turns out that the realistic composite hypothesis representing good lot quality (at most 4% defective) is, for technical reasons, represented effectively by the simple hypothesis (exactly 4% defective), in the sense that the value α characterizing any ('admissible') decision function for the simplified problem is also an upper bound of error probabilities over the realistic composite hypothesis. Similar comments apply to the alternative hypothesis.

[3] The essential point epitomized here is that there is a distinction of levels of language, the first phrase occurring in the 'object language' of things and behavioral acts, the second in the 'metalanguage' in which we discuss a certain statement (hypothesis). Apparent exceptions to the epitomization require explanation in the preceding terms. For example, in a scientific research context 'to decide *that* a certain hypothesis is supported by strong evidence' is tantamount to 'to decide *to* make the statement that the hypothesis is supported by strong evidence.'

The apparently exceptional occurrence here of 'decide to' with an evidential reference is explained by pointing out that 'to make a statement' occurs here in the metalanguage (where all evidential considerations are expressed), and so is not a case of 'to act' when that phrase occurs in the object language, where it has behavioral interpretations.

[4] Examples of joint consideration of these aspects of simple genetics research problems will be found, for example, in Smith (1968) and Mendel (1866). The present writer will offer an extended discussion of such considerations in another paper, 'Mendelian genetics: a case study in the structure of science.'

[5] Even in applications where a behavioral interpretation of 'decisions' clearly applies, the scope of applications of complete formal models of decision problems has had a slow and limited development (see, for example, Brown, 1970); possibly due in part to considerations discussed above.

[6] In the traditional formulation of testing problems the counterpart of the error probability α was the 'probability level' statistic $P = P(x)$. The theoretical aspect of the traditional formulation is a concept of statistical evidence associated with that statistic, under which $P(x)$ is interpreted as an index of strength of evidence against the hypothesis H_1, with smaller values of $P(x)$ indicating stronger evidence. Thus the traditional interpretation is evidential and not behavioral (in any direct sense), and the behavioral interpretation was an innovation of the Neyman–Pearson theory.

In many applications the statistic $P(x)$ was (and is) interpreted schematically, in terms of a dichotomy such as: the statistical evidence against H_1 is strong if and only if $P(x) \leq 0.05$. Here 0.05 corresponds to the error probability α in our schema; and the schematized form of the traditional formulation can be represented by a formal decision function which takes the value 'reject H_1' if and only if the observed sample point x gives $P(x) \leq 0.05$.

[7] This is not to deny that there is any behavioral (literal) realization of certain relative frequencies of errors, approximating the error probabilities in the schema representing a test, in certain long series (actual or conceivable) of applications of tests of the same form. What is suggested is that such a behavioral interpretation is related in a somewhat abstract, indirect (hypothetical or theoretical) way to the evidential interpretation of a single application of a test in a given research situation. This theoretical relation of the evidential meaning of a 'decision' in such an application, to a certain behavioral interpretation of the same formal 'decision' in another context (a series of applications), does not reduce or eliminate evidential interpretations in favor of behavioral ones. On the contrary, appreciation of such a behavioral interpretation, coupled with appreciation of the hypothetical theoretical relation it bears to an evidential interpretation in the given research situation, may be regarded as an important part of appreciation of the meaning of statistical evidence as interpreted under the confidence concept.

[8] The likelihood approach (Edwards, 1972) is based on a primitive concept of statistical evidence which appears closely analogous to our formulation (Conf) of the confidence concept, but which nevertheless does not satisfy the latter nor provide the kind of theoretical control of error probabilities mentioned above. It was rejected by Neyman and Pearson in favor of the confidence concept in their 1933 paper, after they had used it as the basis of their exploratory 1928 paper. A detailed discussion of incompatibilities between the two concepts is given in Birnbaum (1969).

The likelihood concept may be formulated thus:

(L'): If an observed sample point has very small probability (density) under H_1, relative to its probability (density) under H_2, then it provides strong statistical evidence for H_2 as against H_1.

The likelihood and confidence concepts were taken up successively by Neyman and Pearson as plausible successors to the simpler primitive concept of statistical evidence which has been associated (usually implicitly) with tests in their traditional formulation, which has been represented in applications since 1710. Both (Conf) and (L') may be considered as assimilating, in analogous ways, that traditional concept, which may be formulated thus:

(P): A concept of statistical evidence is not plausible unless it finds 'strong evidence against H_1' with very small probability when H_1 is true.

In traditional practice this concept had been complemented by unformalized judgement exercized in the devising and selection of test statistics, which were then interpreted as indices of strength of statistical evidence against a hypothesis H_1, without explicit reference to alternative hypotheses.

Each of the concepts of evidence mentioned may be regarded as a refined version of that simpler familiar intuitive concept which moves us, when something observed seems 'improbable' or 'unlikely' (in any sense, often not specified explicitly), toward reconsideration of some hypothesis, perhaps only tacitly held.

[9] The reader is urged to compare this discussion with the original versions of the argument by Savage and Lindley cited in Section 8.

BIBLIOGRAPHY

Barndorff–Nielsen, O., 1971, *On Conditional Statistical Inference* (mimeographed), Aarhus.

Birnbaum, A., 1961, 'Confidence Curves: An Omnibus Technique for Estimation and Testing Statistical Hypotheses', *Journal of the American Statistical Association* **56** (1961), 246–249.

Birnbaum, A., 1962, 'On the Foundations of Statistical Inference', *Journal of the American Statistical Association* **57** (1962), 269–326 (with discussion).

Birnbaum, A., 1969, 'Concepts of Statistical Evidence', in *Philosophy Science, and Method: Essays in Honor of Ernest Nagel* (ed. by Sidney Morgenbesser, Patrick Suppes, and Morton White), St. Martin's Press, New York.

Birnbaum, A., 1970, 'On Durbin's Modified Principle of Conditionality', *Journal of the American Statistical Association* **65** (1970), 402–403.

Birnbaum, A., 1972a, 'The Random Phenotype Concept, with Applications', *Genetics* **72** (1972), 739–758.

Birnbaum, A., 1972b, 'More on Concepts of Statistical Evidence', *Journal of the American Statistical Association* **67** (1972), 858–861.

Birnbaum, A. and Maxwell, A. E., 1960, 'Classification Procedures Based on Bayes Formula', *Applied Statistics* **9** (1960), 152–159.

Braithwaite, R. B., 1954, *Scientific Explanation*, Cambridge University Press.

Brown, R. V., 1970, 'Do Managers Find Decision Theory Useful?', *Harvard Business Review*, May–June.

Buehler, R. J., 1959, 'Some Validity Criteria for Statistical Inference', *Annals of Mathematical Statistics* **30** (1959), 845–863.

Buehler, R. J. and Fedderson, A. P., 1963, 'Note on a Conditional Property of Student's t', *Ann. Math. Statist.* **34** (1963), 1098–1100.

Cox, D. R., 1958, 'Some Problems Connected with Statistical Inference', *Annals of Mathematical Statistics* **29** (1958), 357–372.

Cox, D. R., 1971, 'The Choice Between Alternative Ancillary Statistics', *Journal of the Royal Statistical Society* **33** (B) (1971), 251–255.

Dempster, A. P. and Schatzoff, M., 1965, 'Expected Significance Level as a Sensitivity Index for Test Statistics', *Journal of the American Statistical Association* **60** (1965), 420–436.

Durbin, J., 1970, 'On Birnbaum's Theorem of the Relation Between Sufficiency, Conditionality, and Likelihood, *Journal of the American Statistical Association* **65** (1970), 395–398 (followed by two discussion notes).

Edwards, A. W. F., 1972, *Likelihood*, Cambridge University Press.
Fisher, R. A., 1956, *Statistical Methods and Scientific Inference*, Oliver Boyd, Edinburgh.
Hacking, Ian, 1965, *The Logic of Statistical Inference*, Cambridge University Press.
Jeffreys, H., 1961, *Theory of Probability*, 3rd ed., Oxford University Press, London.
Lindley, D. V., 1971, 'Bayesian Statistics, A Review', Society for Industrial and Applied Mathematics, Philadelphia.
Mellor, Hugh, 1971, *The Matter of Chance*, Cambridge University Press.
Mendel, G., 1866, 'Versuche über Pflanzenhybriden', *Verhandlungen der Naturforschenden Vereins in Brunn* **4** (1865), 3–44. (English translation in *Experiments in Plant Hybridization*, ed. by J. H. Bennett, 1965, Oliver and Boyd.
Morton, N. E., 1955, 'Sequential Tests for the Detection of Linkage', *American Journal of Human Genetics* **7**, (1955), 277–318.
Nagel, Ernest, 1961, *The Structure of Science*, Harcourt-Brace, New York.
Neyman, J., 1938, (2nd ed., 1952), *Lectures and Conferences on Mathematical Statistics*, Washington D.C. Graduate School, U.S. Department of Agriculture.
Neyman, J., 1947, 'Raisonnement inductif ou comportement inductif', *Proceedings of the International Statistical Conference* **3**, 423–433.
Neyman, J., 1957, 'Inductive Behavior as a Basic Concept of Philosophy of Science', *Review of the International Statistical Institute* **25** (1957), 7–22.
Neyman, J., 1962, 'Two Breakthroughs in the Theory of Statistical Decision-Making', *Review of the International Statistical Institute* **30** (1962), 11–27.
Neyman, J. and Pearson, E. S., 1928, 'On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Part I', *Biometrika* **20A** (1928), 175–240.
Neyman, J. and Pearson, E. S., 1933, 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', *Philosophical Transactions of the Royal Society of London* **231** (A), 289–337 (pp. 140–185 in 1967 reprinting).
Neyman, J. and Pearson, E. S., 1936, 'Contributions to the Theory of Testing Statistical Hypothesis', *Statistical Research Memoirs* vol. I. pp. 113–137 (pp. 203–239 in 1967 reprinting.)
Neyman, J. and Pearson, E. S., 1967, *Joint Statistical Papers*, Cambridge University Press.
Pearson, E. S., 1966, *The Selected Papers of E. S. Pearson*, University of California Press, Berkeley, California.
Renwick, J., 1971, 'The Mapping of Human Chromosomes', *Annual Review of Genetics* **5** (1971), 81–120.
Robinson, G. K., 1974, 'Conditional Confidence Properties of Student's t and of the Behrens–Fisher Solution to the Two Means Problem', unpublished.
Savage, Leonard J., 1954, *The Foundations of Statistics*, Wiley, New York.
Savage, Leonard J., 1962, 'Bayesian Statistics', in *Recent Developments in Information and Decision Processes* (ed. by R. E. Machol and P. Gray), Macmillan, N.Y. and London.
Smith, Cedric A. B., 1953, 'The Detection of Linkage in Human Genetics', *Journal of the Royal Statistical Society* **15** (B) (1953), 155–192.
Smith, Cedric A. B., 1959, 'Some Comments on the Statistical Methods Used in Linkage Investigations', *American Journal of Human Genetics* **11** (1959), 289–403.
Smith, Cedric A. B., 1965, 'Personal Probability and Statistical Analysis, with Discussion', *Journal of the Royal Statistical Society* **128** (A) (1965), 469–499.
Smith, Cedric A. B., 1968, 'Linkage Scores and Corrections in Simple Two- and Three-Generation Families', *Annals of Human Genetics* **33** (1968), 127–150.

Stone, M., 1969, 'The Role of Significance Testing: Some Data with a Message', *Biometrika* **56** (1969), 485–493.
Tukey, J. W., 1960, 'Conclusions v. Decisions', *Technometrics* **2** (1969), 423–433.
Wald, A., 1950, *Statistical Decision Functions*, Wiley, New York.