# Philosophy of Science

December, 1985

## BEHAVIORISTIC, EVIDENTIALIST, AND LEARNING MODELS OF STATISTICAL TESTING*

DEBORAH G. MAYO†

*Department of Philosophy*
*Virginia Polytechnic Institute and State University*

While orthodox (Neyman-Pearson) statistical tests enjoy widespread use in science, the philosophical controversy over their appropriateness for obtaining scientific knowledge remains unresolved. I shall suggest an explanation and a resolution of this controversy. The source of the controversy, I argue, is that orthodox tests are typically interpreted as rules for making *optimal decisions* as to how to *behave*—where optimality is measured by the frequency of errors the test would commit in a long series of trials. Most philosophers of statistics, however, view the task of statistical methods as providing appropriate measures of the *evidential-strength* that data affords hypotheses. Since tests appropriate for the behavioral-decision task fail to provide measures of evidential-strength, philosophers of statistics claim the use of orthodox tests in science is misleading and unjustified. What critics of orthodox tests overlook, I argue, is that the primary function of statistical tests in science is neither to decide how to behave nor to assign measures of evidential strength to hypotheses. Rather, tests provide a tool for using incomplete data to *learn* about the process that generated it. This they do, I show, by providing a *standard* for distinguishing differences (between observed and hypothesized results) due to accidental or trivial errors from those due to systematic or substantively important discrepancies. I propose a reinterpretation of a commonly used orthodox test to make this *learning model* of tests explicit.

With the growing emphasis on the behavioral and social sciences . . . and given the great dependence of these sciences upon statistical methods one must take seriously the claim, from respectable quarters,

*Received August 1984; revised October 1984.

†I am grateful to Ronald Giere, Norman Gilinsky, I. J. Good, Oscar Kempthorne, Henry Kyburg, and Larry Laudan for very helpful comments. I thank Jim Fetzer for first suggesting I spell out my (learning) model by contrasting it to the existing (behavioristic and evidentialist) models of statistical tests.

*Philosophy of Science*, 52 (1985) pp. 493–516.

that the statistical methods currently employed are fundamentally misconceived. R. N. Giere 1969, p. 372.

**1. Introduction and Summary.** In the sixteen years that have passed since this passage from Giere appeared, there is no doubt that philosophers of statistics have taken the earlier criticisms of *orthodox* (or Neyman-Pearson) statistical tests seriously. But far from offering defenses for the widespread use of orthodox tests in science, most, like Rosenkrantz (1977, p. 221), have further "stressed the failure of orthodox theory to provide a satisfactory format for objective scientific reporting (or for conveying 'what the data have to tell us')." Fetzer (1981), Kyburg (1971) and (1974), Levi (1980), Seidenfeld (1979), Spielman (1973), and others have also offered arguments to *strengthen* already existing criticisms of the appropriateness of orthodox tests—at least so far as they are able to perform the task of statistical inference in science.

Such criticisms of orthodox tests arise from opposing views of the appropriate role of statistical tests in science. The views of the major disputants in the testing controversy fall roughly into two camps. I will refer to these as the *behavioral-decision* (or behavioralist) *view,* and the *evidential-strength* (or evidentialist) *view*.

The statistical philosophy of the first camp holds that when evidence is inconclusive all talk of "inferences" and "reaching conclusions" should be abandoned. Rather, the task of a theory of statistics is to provide rules that help guide our behavior with respect to uncertain phenomena, so that we will avoid making erroneous decisions too often in the long run of experience. Accordingly, tests are interpreted as rules of *inductive behavior* yielding the *behavioristic model* of tests, typically associated with Neyman and Pearson.

On the evidential-strength view, on the other hand, when evidence is inconclusive what is needed is some way of quantitatively assessing the extent of the evidence that *particular* observations afford hypotheses. On this view, the task of a theory of statistical inference is to provide an appropriate measure of evidential-relationship, which I abbreviate as an *E-R measure*. Examples include measures of degrees of support, belief, confirmation, corroboration, probability, and the like. Orthodox tests, whose only quantities are long-run error rates of *procedures,* will not be judged adequate for this task unless these error rates can be construed as providing appropriate E-R measures. Attempts at such "evidentialist" interpretations of orthodox tests give rise to what I call *evidential-strength models* of tests.

The problem that arises is this: If orthodox tests are interpreted and judged along the lines of the behavioristic model, then the tests appear appropriate for routine decision-theoretic tasks, where the main concern

is with low long-run frequency of error. But then the tests appear inappropriate for the task of scientific inference. On the other hand, if the error rates of orthodox tests (e.g., significance levels) are interpreted as providing E-R measures (in an attempt to render tests relevant for scientific inference) tests lead to misleading and even contradictory conclusions. So, orthodox tests, if interpreted behavioristically, are inappropriate for scientific inference, and, if interpreted "evidentially," are misleading and contradictory—or so the critics allege. The general thrust of these criticisms is well captured in a passage from Fetzer (1981, p. 244):

> The "preference procedures" Neyman and Pearson have proposed, in other words, may be perfectly suitable for decision-making between restricted alternatives without also fulfilling the appropriate conditions for drawing inferences on the basis of empirical evidence.

Although I take these criticisms of orthodox tests seriously, I deny that they vitiate the manner in which tests can (and very often do) serve their most important function in scientific inquiry. For, what these criticisms of tests overlook, I claim, is that the primary function of statistical tests in science is neither to decide how to behave nor to assign measures of evidential strength to hypotheses. Rather, their primary function seems to come closer to the view expressed by Kempthorne (1971, p. 492) in characterizing statistical inference "loosely as the collection of processes by which we learn from data," as well as the view of E. S. Pearson (1955, p. 204).

While the aim of learning from incomplete data is implicit in much of actual statistical practice, it will not be possible to defend these uses of tests against the well-known criticisms until the manner in which tests serve this distinct function is made explicit. My aim in this paper is to propose a reinterpretation of a commonly used orthodox test in order to make this learning function precise.

To this end, I shall do the following: First, I shall explain, keeping mathematical details to a minimum, enough of the properties of orthodox tests so that the criticisms and proposed resolutions may be understood. Next, I shall show how these criticisms arise from both the behavioristic and the evidential-strength models of tests. Thirdly, I propose a model of tests, which, while retaining the key properties of orthodox tests, is neither behavioristic nor evidentialist (in the sense being used here). To distinguish it from these two other models of testing, I shall refer to this new interpretation as the *learning model* of tests. On the criterion for a "good test" that emerges, I argue, the orthodox tests are appropriate for scientific learning.

The real importance of introducing something like the learning model

of tests into the philosophical discussions of statistical tests, however, goes beyond the desire to give a new defense for orthodox statistics. With a model of tests that accords with actual statistical practice, a new avenue is opened, I believe, for exploring key methodological strategies based on statistical principles of generating and analyzing data.

**2. Orthodox Statistical Tests: Basic Properties.** It should be noted at the outset that the orthodox theory of statistical tests does not provide a single method of testing, but rather a whole conglomeration of different types of tests. A single scientific inquiry typically requires the use of numerous different tests; each directed at answering a different question. For the present purpose, however, it will suffice to consider a type of scientific question that leads to a very commonly used statistical test. Imagine an inquiry into a population of items, say a certain species of fish. A question that may be posed is whether this population of fish differs from some other population of fish, say in being longer (possibly one is interested in the effects of some new fish food, or wants to identify the species).[1] Suppose the question concerns the *average length* of fish in the population being studied, which we symbolize as parameter $\theta$. One hypothesis, let us say, $\mathcal{H}$, is that $\theta$ equals 12 inches; another, $\mathcal{J}$, is that $\theta$ exceeds 12 inches by some unspecified amount. We want to make some observations to test these claims. Ignoring for now the problems of experimental design,[2] a sample of $n$ fish are observed and their lengths appropriately measured, say, at the longest point. A fish's length in inches may be represented by variable $X$; that is, to each fish a value of $X$ (like a little badge) is attached. There are two cases where our observations would give conclusive answers to questions about the average length $\theta$: (1) The lengths of fish do *not vary* at all (for then observing a single $X$-value tells us the population average); or (2) The *entire population* of fish is observed and measured (for then the observed average *is* the population average). More realistically, the values of $X$ are not constant, but are known to vary (among fish in the population), and typically the most we

[1] I deliberately divorce this illustration from any of the possible uses to which such an inquiry may be put so as to concentrate on the general interpretation of tests I will propose. A fuller discussion of an application of tests as well as more of the mathematical details occurs in Mayo (1983). A good account of orthodox tests in general occurs in Kempthorne and Folks (1971).

[2] I mean only that I will not explicitly discuss problems of experimental design here, not that the present treatment is inapplicable to those problems. In fact such problems can usually be dealt with by asking questions about whether certain test assumptions (e.g., independence, control of extraneous variables) are approximately met; and these questions can also be dealt with by means of orthodox tests. Thus, if we can give an adequate account of orthodox tests, we will also be giving an adequate account of a tool needed for dealing with experimental design problems.

can observe is some proper subset of the population. In these cases statistical considerations are needed for testing claims about $\theta$.

Suppose it is known that values of $X$ vary according to a pattern closely resembling that of a Normal distribution with an average (mean) value-$\theta$ (which is in question) and a known standard deviation $\sigma$ of 2. Having observed the lengths of the $n$ fish in our sample, the most useful statistic to calculate is the average (mean) length, denoted by $\bar{X}$ in the sample; for it varies least, on the average, from the population parameter $\theta$ of interest.[3] But even if $\mathcal{H}$ is true and our sample does come from a population of fish whose average length $\theta$ is 12, it does not follow that the observed average in this sample will be exactly 12. What follows is only that the most frequent observed outcome is expected to be 12 and that small differences from 12 will be more frequent than differences far from 12. This prediction can be expressed as the statistical hypothesis H: $\bar{X}$ follows the Normal distribution with mean value $\theta$ equal to 12, and standard deviation $\sigma_{\bar{X}}$ (which equals $\sigma/n^{1/2}$). A population of larger fish, on the other hand, is associated with a $\theta$ that exceeds 12. The sort of statistical test frequently used in such an inquiry involves the following *null* and *alternative* hypotheses:

(2.0) *Null Hypothesis H:* $\bar{X}$ is Normal ($\theta$, $\sigma_{\bar{X}}$) and $\theta = 12$
*Alternative Hypothesis J:* $X$ is Normal ($\theta$, $\sigma_{\bar{X}}$): and $\theta > 12$

where the standard deviation $\sigma_{\bar{X}} = 2/n^{1/2}$. The null hypothesis is *simple,* in that it specifies a single value of $\theta$, while the alternative is *composite,* as it consists of the set of $\theta$-values exceeding 12.

Since our test is devised so as to reject $H$ just in case $\theta$ exceeds 12, i.e., our test is *one sided* (in the positive direction), it seems plausible to reject $H$ on the basis of sample averages ($\bar{X}$ values) that exceed 12 sufficiently; and this is precisely what the orthodox test recommends. That is, our *test rule,* which we may represent by $T^+$, rejects $H$ just in case $\bar{X}$ is "significantly far" (in the positive direction) from hypothesized average 12, where *distance* is measured in standard deviation units (i.e., in $\sigma_{\bar{X}}$'s).

Denoting the *observed average* by $\bar{X}_{\text{obs}}$, the *average hypothesized* by $H$ by $\theta_H$, we can abbreviate its *observed distance* (for $\theta_H$) by $D_{\text{obs}}$ where

[3]A standard measure of the average deviation of $\bar{X}$ from $\theta$ is the *standard deviation* of $\bar{X}$, denoted by $\sigma_{\bar{X}}$. If $X$ follows the normal distribution with mean $\theta$ and standard deviation $\sigma$, then $\bar{X}$ also follows the normal distribution with mean $\theta$, only now its standard deviation $\sigma_{\bar{X}}$ equals $\sigma$ divided by the square root of sample size $n$; i.e., $\bar{X}$ is Normal ($\theta$, $\sigma/n^{1/2}$).

What makes statistic $\bar{X}$ so valuable is that its distribution is (approximately) Normal ($\theta$, $\sigma/n^{1/2}$) with $\theta$, $\sigma$ equal to the mean and standard deviation of the underlying distribution of $X$, respectively, *no matter what* this underlying distribution is (barring an infinite $\sigma$). This is the essence of the *Central-Limit Theorem.*

(2.1) $D_{\text{obs}}$ = (Observed Average ($\bar{X}_{\text{obs}}$)) − (Hypothesized Average ($\theta_H$) [e.g., 12])

equals some number $k$ of standard deviation units, (i.e., $D_{\text{obs}} = k\ \sigma_{\bar{X}}$). Corresponding to each such observed difference (naturally, $D_{\text{obs}}$ varies as $\bar{X}_{\text{obs}}$ does) is its *level of statistical significance*.

(2.2) *The Statistical Significance Level of an observed difference* $D_{\text{obs}}$ (in testing $H$) equals the frequency with which so large a difference arises assuming $H$ is true.[4]

An orthodox test consists of a rule which specifies, *before* the observation is made, how statistically significant an observed difference must be before it should be taken to reject $H$. The maximum significance level chosen beyond which $D_{\text{obs}}$ is taken to reject $H$ is called the *size* of the test, and is denoted by $\alpha$. In the case of $T^+$ we have

(2.3) *Test Rule* $T^+$ *with size* $\alpha$: Reject $H$ at level $\alpha$ iff $D_{\text{obs}}$ is statistically significant at level $\alpha$ (i.e., iff so large an observed difference occurs no more than $\alpha$(100 percent) of the time if $H$ is true).

The smaller the value of $\alpha$ chosen as the size of the test, the less frequently level $\alpha$ is reached (and so $H$ is rejected) when in fact $H$ is true. More specifically, a test with size $\alpha$ rejects $H$ when in fact $H$ is true (i.e., it commits a *Type I error*) no more than $\alpha$(100 percent) of the time. That is,

(2.4) The *size* of a test equals the frequency with which the test erroneously rejects $H$ (in a sequence of applications of the test rule).

For, a test with size $\alpha$ rejects $H$ just in case $D_{\text{obs}}$ reaches the $\alpha$ level of significance. But by definition (see (2.3)) this occurs no more than $\alpha$(100 percent) of the time when $H$ is true. In testing our null hypothesis $H$:$\theta = 12$ (against $J$:$\theta > 12$), the following test rules have sizes .02 and .001, respectively:

[4]By "so large a difference" I mean one *as large as or larger than* the observed difference. Suppose the average length in a sample of 25 fish is observed to be 12.1 inches, ($\bar{X}_{\text{obs}} = 12.1$). The difference between the observed and hypothesized averages, taking $\theta_H$ to be 12, is 12.1 minus 12, giving an observed difference of .1 ($D_{\text{obs}} = .1$).

Since $\sigma$ (according to $H$) equals 2, and $n$ equals 25, $\sigma_{\bar{X}}$ equals $^2/_5$ or .4 inches. So, $D_{\text{obs}} = .1 = {}^1/_4\ \sigma_{\bar{X}}$. That is, if $H$ is true, and $\theta$ equals 12, then the observed-sample average differs from the population average by $^1/_4\sigma_{\bar{X}}$'s. We can ask: How frequently does an observed-sample average exceed its population average by *as much as or more than* $^1/_4$ standard deviation units? This is identical to asking: How *statistically significant* is a difference of $^1/_4\sigma_{\bar{X}}$? The answer turns out to be .4; that is, such a large observed difference occurs 40 percent of the time when observing a population correctly described by $H$. This would not be considered good grounds for rejecting $H$. Note, in contrast that 12.1 *is* significant at the .02-level in test $T^+$-1600; it represents a difference of $2\sigma_{\bar{X}}$ (each $\sigma_{\bar{X}}$ now being $^2/_{40}$ or .05).

(2.5) (a) $T^+$: Reject $H$:$\theta = 12$ at level .02 iff $D_{\text{obs}} \geq 2\sigma_{\bar{X}}$.
(b) $T^+$: Reject $H$ at level .001 iff $D_{\text{obs}} \geq 3\ \sigma_{\bar{X}}$.

Clearly then, an observation $\bar{X}_{\text{obs}}$ may lead to rejecting $H$ with rule (a) and not with rule (b); (for (b) requires the observed difference to be larger than (a) does before it rejects $H$).[5] As such, whether or not a test leads to rejecting $H$ is a function of how the size $\alpha$ is specified. But how is the size of a test to be chosen?

Since the smaller the $\alpha$ the less frequent the Type I error, it may be thought that $\alpha$ should be made as small as possible. But by making $\alpha$ smaller (and so making a rejection of $H$ more difficult), the test suffers an increase in the frequency with which it *fails to reject H* (i.e., accepts $H$) even when in fact $H$ is false (and so *should* be rejected). An *erroneous acceptance* of $H$ is called the *Type II error,* and the frequency of a Type II error is denoted by $\beta$. $\alpha$ and $\beta$ are the *error frequencies* (or error probabilities) of tests:

(2.6) $f(T^+$ Rejects $H|H$ is true$) \leq \alpha$ = frequency of Type I error
$f(T^+$ Accepts $H|J$ is true$) \leq \beta$ = frequency of Type II error.

(In an extreme case where $\alpha$ is set at 0, the test never erroneously rejects $H$, since it never rejects $H$ altogether. But, if $H$ *is* false, such a test will always *accept H* erroneously, i.e., $\beta = 1$.) It should be noted that with a composite alternative, as in example $T^+$, the value of $\beta$ varies with different alternative values for $\theta$, i.e., it varies according to "how false" $\theta_H$ is. The more discrepant $\theta$ is from $\theta_H$, the less frequent an erroneous acceptance of $H$ occurs, i.e., the smaller is the value of $\beta$. Although introducing the second type of error helps to constrain the specifications of a test's error rates, there are still numerous ways of balancing $\alpha$ with $\beta$.

The task of specifying the error rates of tests is considered to lie outside the domain of the formalism of orthodox tests, and this has resulted in the tests being criticized as lacking in objectivity.[6] But there is an important sense in which orthodox tests are objective; namely, they guarantee that the frequency of errors will not exceed the error rates one spec-

[5]Rules (2.5) (a) and (b) can equivalently be written:

(a) Reject $H$ at level .02 iff $\bar{X}_{\text{obs}} \geq \theta_H + 2\ \sigma_{\bar{X}}$ and
(b) Reject $H$ at level .001 iff $\bar{X}_{\text{obs}} \geq \theta_H + 3\ \sigma_{\bar{X}}$.

Let $\bar{X}_{\text{obs}} = 12.8$, $n = 25$, $\theta_H = 12$. Since $\sigma_{\bar{X}}$ is then .4, our observed difference equals 2 $\sigma_{\bar{X}}$ exactly. So rule (a) maps this difference to "Reject $H$" while (b) does not. $\bar{X}_{\text{obs}}$ would have to be at least 13.2 before (b) led to rejecting $H$.

[6]In Mayo (1983) I argue that these criticisms are based on an overly narrow conception of objectivity and of what is required for objective learning in science. I defend an altered conception of objective learning and argue that orthodox tests may be reformulated so as to serve as a means for objective learning in science.

ifies the test to have. For, an orthodox test, considered in its naked mathematical form alone, is essentially this:

(2.7) An *orthodox test* is a rule that maps each of the possible values observed into either Reject $H$ (Accept $J$) or Accept $H$ in such a way that it is possible to guarantee *before the trial* is made, that (regardless of the true value of $\theta$) the rule will erroneously reject $H$ and erroneously accept $H$ no more than $\alpha$(100 percent) and $\beta$(100 percent) of the time, respectively.

Yet, "from the point of view of mathematical theory all that we can do is show how the risk of the errors may be controlled and minimized" (Neyman and Pearson 1933, p. 146). Whether or not such a piece of mathematics is appropriate for the task of statistical tests in science depends both on one's view of this task and on the interpretation with which one clothes the formal components of tests.

**3. The Behavioristic Model of Tests.** In order to provide some objective basis for specifying and interpreting the tests that were already being used in science in the 1920s (particularly those of R. A. Fisher), Neyman and Pearson consider a paradigm type of context in which such an objective basis would be forthcoming. Here it is imagined that one can specify how often one "can afford" to make the Type I and Type II errors by appealing to the "seriousness" of the consequences of doing so.

But when are considerations of how often one "can afford" to be wrong, and of the "seriousness" of certain errors, forthcoming? Noting that such considerations arise in certain decision-theoretic contexts, Neyman and Pearson are led to suggest the *behavioristic* model of tests. Although the behavioristic construal of tests was largely advocated by Neyman, and was not wholly embraced by Pearson (see Pearson 1955), the Neyman and Pearson theory of tests (and often orthodox tests as a whole) is generally viewed as providing a model of *inductive behavior*.

Here, tests are formulated as mechanical rules, or "recipes" for reaching one of two possible decisions: "accept hypothesis $H$" or "reject $H$," where these are interpreted as deciding to "act as if $H$ were true" and "act as if $H$ were false," respectively.

> Here, for example, would be such a 'rule of behavior': to decide whether a hypothesis $H$, of a given type, be rejected or not, calculate a specified character, $x$, of the observed facts; if $x > x_0$ reject $H$; if $x \leq x_0$, accept $H$. Such a rule tells us nothing as to whether in a particular case $H$ is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule . . . we shall reject $H$ when it is true not more, say, than once in a hundred

> times, and in addition we may have evidence that we shall reject $H$ sufficiently often when it is false.(Neyman and Pearson 1933, p. 142)

The view that emerges is this: since one cannot infer the truth of a hypothesis on the basis of incomplete data, statistics should provide, not rules of inductive inference, but rules for making optimal decisions as to how to behave with respect to hypotheses. And a theory of statistics can perform this function by providing rules that assure that one would not be wrong too often. The criterion for a "good test" on the behavioristic model (abbreviated as [BM]) may be stated thus:

(3.0) [BM]: A *good test* is one that has an appropriately small frequency of rejecting $H$ erroneously, and at the same time erroneously accepts $H$ sufficiently infrequently (in a given sequence of applications of the rule).

Can the formal apparatus of orthodox tests satisfy the criterion of [BM]?

The statement arrived at in (2.7) makes it plain that the answer is yes. For the formal apparatus of an orthodox test guarantees that the test's error rates will not exceed the values of $\alpha$ and $\beta$ that one selects; one needs only to fix them at appropriately small values. Neyman and Pearson propose that one first fix $\alpha$ at some suitably small value, and then seek the test which at the same time has a suitably small $\beta$. The "best" test of a given size $\alpha$ (if it exists) is the one that at the same time minimizes the value of $\beta$ (i.e., the rate of Type II errors) for all possible values of $\theta$ under the alternative $J$. And the tests given in 2.5 are the "best" Neyman and Pearson tests $T^+$ with sizes .02 and .001, respectively.

However, are tests that are "good" according to behavioristic criteria (of low error-rates in the long run) also good as tools for obtaining scientific knowledge? Is test $T^+$, for example, a good tool for finding out *what is the case,* as opposed to *how to best behave,* with respect to the lengths of a certain species or population of fish? Most philosophers of statistics say no. As Kyburg (1971, p. 82) puts it:

> When it comes to general scientific hypotheses (e.g., that $f(x)$ represents the distribution of weights in a certain species of fish . . .) then the purely pragmatic, decision theoretic approach has nothing to offer us.

The basis for their negative answers is this: It is admitted that if one is in the sort of decision-theoretic context envisioned by the behavioristic approach, then the orthodox test may be sensible. The paradigm example of such a context is acceptance sampling in industrial quality control. But in scientific contexts the behavioral interpretation of accept $H$ and reject $H$ seems out of place. A scientist does not seem to be in a position to specify how often he "can afford" to be wrong in some long run; nor

does the low error-rate in the long-run rationale seem relevant for a scientist who is concerned with what *particular* inference from *this* experiment is warranted. Nevertheless, orthodox tests enjoy widespread use in science.

So it appears that scientists either routinely apply tests that are entirely ill suited to their needs, or else they use orthodox tests in a way that fails to be captured within the behavioristic model found in statistics texts. In reality no statistical consultant worth his or her salt simply sets up an $\alpha$-level test, for a conventionally small $\alpha$(.01 or .05), and then rejects or accepts $H$ according to whether or not the observation is significant at level $\alpha$. From the start, Pearson (1947, p. 192) declared that "no responsible statistician, faced with an investigation of this [non-routine] character, would follow an automatic probability rule." But, the still unanswered question is: How *are* statistical tests to be used in scientific inquiry?

**4. Evidential-Strength Models of Statistical Tests.** Existing attempts to answer this question have endorsed what I refer to as *evidential-strength interpretations* or *models* of tests.[7] Such attempts arise out of an assumption that has been unquestionably accepted in discussions of philosophy of statistics; namely, that the task of a theory of statistics (in science) is to provide some means of using data to assign hypotheses a measure of evidential strength (support, probability, reliability, degree of belief, etc.). Carnap, Hacking, Kyburg, Levi, Salmon, Seidenfeld, and others have endorsed one or another such measures of evidential relationship between data and hypotheses. I will abbreviate all such measures as E-R measures. Our purpose here is not to evaluate their separate systems (but see Mayo 1981b) but to consider the evidential-strength model of orthodox tests that arises from this tradition, and to explain in a very general way why orthodox tests fail to satisfy the testing criterion of the evidential-strength model.[8]

[7]Myself and Giere (e.g., Giere [1976]) are exceptions but see n. 16 for some differences. Kempthorne and Folks (1971) also suggest a reinterpretation of orthodox tests in terms of a notion [consonance] that is something other than an evidential-strength measure. It should be noted that evidential-strength models of tests typically arise within attempts to criticize, rather than defend tests.

Two notable exceptions involve attempts to erect plausible E-R concepts that are based on orthodox testing ideas (i.e., error probabilities), but which avoid certain "evidentialist" criticisms. The first is the work of Birnbaum (1977), who, unfortunately, died before fully explicating his notion of a *confidence concept*. The second is the *standardized tail-area* notion developed by Good (1982) as a way of using significance levels evidentially, while avoiding criticisms based on (4.1). (See n. 9.)

[8]Mayo (1982) provides a specific explanation of why Neyman-Pearson tests fail to satisfy the evidential-strength criteria by which, I claim, tests are judged by Hacking, Seidenfeld, and Spielman. However, the failure of Neyman-Pearson tests to satisfy these criteria, I argue, can only be taken as grounds for criticizing these tests by misinterpreting the tests, or, by begging the question against them.

On the evidential-strength view the orthodox tests would appear adequate only if "accept $H$ and reject $H$ (with test $T$)" could be construed as something like "strong evidence in favor of $H$" and "strong evidence against $H$," respectively, where evidential strength is measured by a chosen E-R measure. That is, the criterion for a "good test" on the evidential-strength model (abbreviated [EM]) is this:

(4.0) [EM]: A good test rejects $H$ (accepts $J$) iff observed data (e.g., $\bar{X}_{obs}$) provides appropriately strong evidence against $H$ and in favor of alternative $J$ (i.e., weak evidence in favor of $H$ as against $J$).

Is an orthodox test that is "good" according to the criterion of the behavioristic model (i.e., [BM]) also "good" according to the criterion of the evidential-strength model [EM]? One apparent way of getting a yes answer to this question is by interpreting "small frequency of erroneously rejecting $H$" (or, "low significance level $\alpha$") as something like "small evidential support for $H$ (as opposed to $J$)"; in short, by interpreting orthodox error-probabilities as measures of evidential strength (i.e., as E-R measures). Then the (pragmatic) decisions of the behavioral model become construed as (cognitive) decisions to assign $H$ one or another E-R measure. So it appears at first glance that our problem is solved by an evidential-strength interpretation of orthodox test results.

On such an interpretation, rejecting $H$ at significance level .02, for example, might be interpreted as: assign hypothesis $H$ 2 percent probability, support, or other E-R measure; and assign alternative $J$ 98 percent evidential support. But such interpretations of error frequencies, while common, are unwarranted and conflict with basic principles of orthodox tests (e.g., the frequency view of probability). The only thing a .02-rejection says about *a specific* rejection of $H$ is that it was the result of a *general testing procedure* which erroneously rejects $H$ only 2 percent of the time in the long run of (similar or very different) applications of the test. Since, in an orthodox testing context, parameter $\theta$ is viewed as a fixed (yet unknown) quantity, hypotheses about it are viewed as either true or false. Thus, it makes no sense (within the orthodox context) to assign them any probabilities other than 0 or 1; that is, a hypothesis is true 100 percent of the time or 0 percent of the time, according to whether it is true or false. But so long as its truth is unknown, the only thing the orthodox test gives us are error frequencies of test rules. And, as the critics of orthodox tests show, an evidential-strength interpretation of error frequencies is unworkable. (But see Birnbaum 1977 for the most promising such attempt.) As Kyburg (1974, p. 58) notes:

> But although many statisticians, and essentially all psychologists, sociologists, political scientists, economists, biologists, ecologists, bacteriologists, pathologists, physicians, toxicologists, astronomers, an-

> thropologists, etc. cite significance levels (the smaller the more proudly) . . . as though they reflected a *level of evidential support applicable to the instance at hand,* we know that in general this cannot be the case. [emphasis added]

Probably the most flagrant and often-cited objection to using significance levels this way is that for a fixed significance level $\alpha$, no matter how small, a large enough sample size can make it overwhelmingly likely that $H$ will be rejected at that level—even on the basis of data which hardly seem to provide evidence against $H$.

One can easily see how this problem arises by referring to our example of test $T^+$. An observed outcome ($\bar{X}_{\text{obs}}$) reaches a significance level of, say, .02, just in case $D_{\text{obs}}$ exceeds 2 $\sigma_{\bar{X}}$'s (see 2.5(a)). And as the sample size $n$ increases, the size of a single standard deviation $\sigma_{\bar{X}}$ decreases (being inversely proportional to $n$). Thus, any difference $D_{\text{obs}}$, as small as one likes, is significantly different (from $H$) at level $\alpha$, for as small an $\alpha$ as one likes, provided $n$ is made sufficiently large. This paves the way for the following sort of "evidentialist" criticism of orthodox tests:

(4.1) (i) Consider a sample mean $\bar{X}$ that is significantly different from $H$: $\theta = \theta_H$ (according to orthodox test $T^+$) at level $\alpha$, for $\alpha$ as small as one wants. (e.g., For $\alpha$ fixed at .02, $\bar{X} = \theta_H + 2\sigma_{\bar{X}}$ would be $\alpha$-significant.)

(ii) For a sufficiently large sample size $n$, such an $\alpha$-significant sample mean will differ so little from hypothesized value $H$ that (on a plausible E-R measure) it is taken as *strong support in favor* of $H$, or at least as little evidence against $H$ (i.e., little evidence in favor of alternatives $J$).

(iii) [From (i) and (ii)]: The "best" test $T^+$ may sanction a rejection of $H$ on the basis of an observation that provides good evidence in favor of $H$ on any plausible E-R measure. Thus, a "good" test according to the [BM] criterion (low error-rates) may fail to be "good" on evidentialist criterion [EM].

(iv) [From (i)–(iii)]: If we interpret "$\bar{X}_{\text{obs}}$ is statistically significant" or "$T^+$ rejects $H$ at *small significance level* $\alpha$ (e.g., .02)" as "$\bar{X}_{\text{obs}}$ provides a *small* amount of *evidential support* for $H$," then (where $n$ is sufficiently large) test $T^+$ may lead to assigning *low support for H* on the basis of data that *highly supports H*.

Statement (iv) may be called the "Jeffreys-Good-Lindley Paradox" after three of the first statisticians to demonstrate it for the case where Bayesian posterior probabilities serve as the E-R measures. The thrust of their arguments, as Lindley (1972, p. 15) summarizes it, is:

> . . . that a result which is conventionally significant at, say, 5%, can

have posterior probability near to 1, so that *a hypothesis can be 'rejected' when it is highly likely to be true*. . . . for large enough $n$ the posterior odds on $\theta_H$ can be as large as one likes for a fixed level of significance.[9]

Numerous criticisms of orthodox tests are variations on the same theme; namely, that tests violate evidentialist criterion [EM] (for one or another choice of E-R measure). Perhaps the most persuasive variety; criticisms of *ultra-sensitive* tests, start by considering a value of $\theta$ deemed *negligibly discrepant* (in the positive direction) from the hypothesized $\theta_H$. Although test $T^+$ includes all positively discrepant $\theta$ values in its alternative $J$ (mainly for ease of mathematics), in reality, one is rarely interested in any and all positive discrepancies from $\theta_H$. Nor is it supposed that a simple (point) null hypothesis $H$: $\theta = \theta_H$ is ever a statement of the precise value of a continuous quantity to any number of decimal places. This leads to an alternative version of argument (4.1):

(4.1)* (i)* Consider a value of $\theta$ (e.g., 12.2), which, while positively discrepant from $\theta_H$ is considered *negligibly discrepant* from $H$. Typically, one would not want to reject $H$ for a given problem or context if $\theta$ exceeds $\theta_H$ by such a trivial amount (e.g., by only .2 inch).

(ii)* By selecting an orthodox test to have an appropriately large sample size $n$ (e.g., 1600) it can be assured that the test almost *always* (e.g., 98 percent of the time) leads to rejecting $H$ at any desired significance level $\alpha$ (e.g., .02) as long as the actual value of $\theta$ exceeds $\theta_H$ by *any amount at all* (e.g., even if $\theta$ is no greater than 12.2).[10]

(iii)* [From (i)* and (ii)*]: Two rejections of $H$, at the same level $\alpha$(e.g., .02) with a given test-rule (e.g., $T^+$) may, if the two tests

[9][I have added underlining for emphasis, and replaced Lindley's $\theta_0$ with $\theta_H$ for consistency with my notation.]

The explanation (for contexts such as $T^+$) is that for sufficiently large sample size $n$, the Bayes factor (an E-R measure) *against* $H$ (and in favor of alternative $J$) is approximately proportional to $1/n^{1/2}$ (so long as the prior probability [density] of parameter $\theta$ is bounded as $\theta$ takes on alternative values approaching $\theta_H$). By choosing a large enough $n$, even an $\alpha$-significant result (were it to occur) would, for a Bayesian, provide little probabilistic support for alternatives to $H$, and so *high support in favor of* $H$. For detailed discussions also see Good (1980, 1981, and 1982); and Edwards, Lindman, and Savage (1963).

[10]From n. 5, (a), we have that a sample mean rejects $H$: $\theta = \theta_H$ at the .02-level just in case it exceeds $\theta_H$ by $2\sigma_{\bar{X}}$ (i.e., $2\sigma/n^{1/2}$). Thus, to ensure $H$: $\theta = 12$ is rejected 98 percent of the time at the .02-level when $\theta$ is as small as, say 12.2, we must ensure a sample mean of $12 + 2\sigma_{\bar{X}}$ will arise 98 percent of the time when $\theta = 12.2$. To do this we set 12.2 equal to $(12 + 2\sigma_{\bar{X}}) + 2\sigma_{\bar{X}} = 12 + 4\sigma_{\bar{X}}$. Equivalently, set discrepancy .2 equal to $4\sigma/n^{1/2}$. Solving for $n$ yields $n = (20\sigma)^2$. Since $\sigma = 2$, we get $n = 1600$.

have different sample sizes, correspond to different degrees of the "falsity" of $H$ (i.e., to different discrepancies between the actual and hypothesized values of $\theta$).

(iv)* [From (i)*–(iii)*]: If we interpret "$T^+$ rejects $H$ at small significance level $\alpha$" as "assign (or decide to assign) low support (or degree of belief, etc.) to $H$," then the "best" test (judged by [BM]) $T^+$ may *almost always* instruct us to assign $H$ low support (or other E-R measure) even though $\theta_H$ is negligibly discrepant from the true value of $\theta$. Moreover, criterion [BM] would not distinguish this test from one in which this would rarely occur (e.g., when $n$ is only 25).

Now a strict follower of the behavioristic model can deny the force of the evidentialist criticisms of tests. For, he could maintain, he is not interested in evaluating data on any of the evidential-strength measures proffered by the critics. Error frequencies are simply long-run error-rates of procedures, and it is not his fault if someone tries to use them as measures of the "goodness" (in the sense of evidential strength) of particular applications of test procedures. Moreover, as Birnbaum (1977) has shown, existing E-R approaches fail to provide the guarantees of error rates in the long run that orthodox tests can. In short, our strict behaviorist retorts, these evidentialist "criticisms" simply emphasize the incompatability between a concern with error rates and a concern with evidential-strength measures.

But the evidentialists want to go further and claim that in order to use results of a statistical test to obtain scientific knowledge, one is forced to adopt a (possibly invalid) evidential-strength interpretation of error frequencies. Lindley (1965, p.68), for example, states:

> . . . the 5% or 1% [significance level] is a measure of *how much belief* is attached to the null hypotheses [$H$: $\theta = \bar{\theta}$]. It is used *as if* 5% significance meant [in terms of Bayesian inference] the posterior probability that $\theta$ is near $\bar{\theta}$ is .05. This is not so: the distortion of the meaning is quite wrong in general. [emphasis added]

Regardless of the choice of E-R measure chosen as appropriate, an essential premise underlying evidentalist criticisms of tests is

> (v) A statistical test is appropriate for scientific knowledge only if it satisfies [EM] (for a suitable choice of evidential-strength measure).

Together with (4.1) and (4.1)*, it is concluded that

> (vi) Orthodox tests are inappropriate for the task of obtaining scientific knowledge.

But why accept premise (v)? So deeply entrenched is the evidential-strength philosophy that (v) (or something like it) has not been thought to require justification. Claims like the following are typical:

> Clearly what is wanted is a continuously variable measure of how probable the various hypotheses are, in the light of the data, and the NPT [Neyman-Pearson test] fails to provide this. One must conclude that it is not an appropriate theory of inference. (Smith 1977, p. 74)

When the strict behaviorist simply denies (v), by asserting that only low error-rates matter, the gulf between orthodox testing principles and such well-entrenched epistemological ones only seems to widen. What is needed is a positive argument showing *how* the formalism of orthodox tests (see (2.7)) may be clothed so as to use test results for "conveying what the data have to tell us." To this end, I propose a third way of interpreting orthodox tests: one that retains the desirable property of orthodox tests (e.g., objective control of error rates), and yet involves a non-behavioralist, non-evidential-strength interpretation of test results.

**5. A Learning Model of Statistical Tests.** On the interpretation I propose, tests are viewed (in scientific inquiries), to use a phrase of E. S. Pearson (1955, p. 204), as providing a "means of learning" from experimental data. Although Pearson did not precisely spell out how he thought tests served this learning function, it seems clear that he held that the value of orthodox tests (in such learning contexts) need *not* lie in the long-run error-rate rationale found in the behavioral model [BM]:

> In other and, no doubt, more numerous cases there is no repetition of the same type of trial or experiment, but all the same we can and many of us do use the same test rules to guide our decision, following the analysis of an isolated set of numerical data. Why do we do this? What are the springs of decision? Is it because *the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgement?* Or is it because we are content that the application of a rule, now in this investigation, now in that, should result in a long-run frequency of errors in judgement which we control at a low figure? (Pearson 1947, p. 173; emphasis added)

Regretfully, Pearson leaves this tantalizing question unanswered, claiming "On this I should not care to dogmatize."

Without being dogmatic, I suggest that in scientific contexts error frequencies are important, not because one is concerned simply with low error-rates in the long run; but because they provide "that clarity of view needed for sound judgment" regarding what has or has not been learned in a given statistical inquiry. I shall refer to this as a *learning model of*

*testing*.[11] Tests accomplish this learning function by providing tools for *detecting* certain *discrepancies* between the (approximately) correct parameter values (within a statistically modeled problem) and the hypothesized ones. In our example of $T^+$, one is interested in learning the extent to which the actual value of $\theta$ is positively discrepant from the hypothesized value, namely 12 (see 4.1*). Suppose our statistical result is an $\alpha$-level rejection of $H$ with test $T^+$. How can error probabilities be used to determine the discrepancies about which we have and have not learned? An analogy may be helpful.

A test of type $T^+$ may be construed as an instrument for categorizing observed-sample averages according to the size of the mesh of a netting on which they are "caught." In our example we are interested in "catching" fish larger than those arising from a population of fish correctly described by $\theta = 12$. A test categorizes a sample average as "significant at level $\alpha$" just in case it is caught on a size $\alpha$ net; where a size $\alpha$ net is one that would catch $\alpha$(100 percent) of the possible samples from a fish population where $\theta = 12$. We imagine that if a sample average ($\bar{X}_{\text{obs}}$) is "caught" on a size $\alpha$ net, then it would fall through any larger sized mesh; i.e., it is not statistically significant at any smaller size $\alpha$. The smaller the value of $\alpha$ chosen for the $\alpha$-significant net, the larger the width of the netting on which the observed average must be caught—so the smaller the percentage of the fish that are caught on it.

Then specifying a test to have a small size $\alpha$ is analogous to rejecting $H$: $\theta = \theta_H$ (e.g., 12) just in case a given sample of fish has an average length ($\bar{X}_{\text{obs}}$) that is "caught" on a net on which only a small percentage of samples from fish population $H$ would be caught. The rationale for a reasonably small $\alpha$ is this:

(5.0) If it would be *very rare* for so large a catch to arise in a population of fish with average length no greater than $\theta'$, then such a catch is a *good indication* that one is fishing from a population where $\theta$ exceeds this value $\theta'$.

Although this justifies taking an $\alpha$-rejection of $H$ (for small $\alpha$) as indicating that *some* positive discrepancy between $\theta$ and $\theta_H$ has been detected; it is still not clear how the problem raised in (4.1)* is to be avoided. For that argument shows that a rejection of $H$ with test $T^+$ even at a small

[11]If an evidential interpretation is simply seen as one in which data are used to reach true claims (about what the data convey about a statistically modeled problem) and avoid false ones (i.e., avoid misinterpreting the data), then the interpretation I propose does constitute an "evidential" interpretation. However, that term has been so closely tied to the view that a theory of statistics must provide an assessment of the evidential strength that data afford hypotheses, that it seems clearer to designate the present interpretation by means of a different term altogether.

$\alpha$ level, such as .02, may reflect far less of an underlying discrepancy when it arises from a test with large sample size, say $n = 1600$, as it does from a test with a much smaller sample size, say $n = 25$. Let us abbreviate these two $T^+$ tests by $T^+$-1600 and $T^+$-25, respectively. But, as noted in (4.1)*(iii)*, the test criterion [BM] distinguishes tests only by their long-run error-rates (i.e., their "operating characteristics"). So the report: "Reject $H$ at level .02 with the best test of type $T^+$," when it arose from $T^+$-1600 would not be distinguished from its having arisen from $T^+$-25. Clearly then, if one wants to distinguish the discrepancies indicated by the two .02-rejections of $H$, one must go beyond the criterion of the behavioral model of tests. To illustrate how the learning model does this, consider again the fishnet analogy of tests:

Imagine that two fisherman, Mr. Powers and Mr. Coarse, seek lakes with fish that are longer on the average than those in the lake in which they usually fish, call the latter Null Lake. Mr. Powers tries fishing in Lake A, Mr. Coarse, in Lake B, where for convenience they use nets rather than rods. At the end of the day each fisherman claims to have netted fish significantly larger (as measured by their average length $\bar{X}$) than what they would typically have netted in Null Lake, whose fish average only 12 inches.

> *Mr. Powers:* If I had been netting in Null Lake, I'd have gotten a catch as large as today's catch only two times out of 100.
>
> *Mr. Coarse:* Ditto.

Suppose it turns out that the size of the netting Mr. Powers used is far smaller than the size of Mr. Coarse's netting. Then, we would rightfully conclude that Mr. Powers had detected less of a discrepancy between the size of fish in Lake A and those of Null Lake, than Mr. Coarse found in Lake B. More specifically, suppose we find out the following. Using Mr. Coarse's net, not only is today's catch a very rare (frequency .02) occurrence when fishing in Lake Null (where $\theta = 12$ inches), it is also quite *a rare occurrence* (frequency .06) for a lake with fish averaging 12.2 inches. Then, following the principle in (5.0), Mr. Coarse's catch *would* indicate he was fishing in a lake where $\theta$ exceeded 12.2.

Suppose, on the other hand, that Mr. Powers's net is far more sensitive than Mr. Coarse's. Although it is true that Mr. Powers's catch (or one even larger) would arise very rarely (only 2 percent of the time) if he had been netting in Lake Null (and so by (5.0) *some* discrepancy from Lake Null is indicated); say it is also true that "such a catch" (i.e., one as large or larger) *occurs very frequently,* (in fact 98 percent of the time) from a population of fish with average length $\theta$ only 12.2. In other words, Mr. Powers could be expected to be as or even more excited than he is about today's catch (using his net) 98 percent of the time, even if he were

fishing in a lake whose fish averaged only 12.2 inches! If he maintained that his catch indicated an average fish size $\theta$ in excess of 12.2, we would consider him to be misconstruing his results (i.e., making great whales out of little flounders.) The principle that emerges is this:

(5.1) If a given catch would arise fairly frequently from a population with average fish length $\theta$ no greater than $\theta'$ then such a catch does *not* indicate that one is fishing from a population where $\theta$ exceeds $\theta'$.

This reasoning makes it clear why one would deem the result of Mr. Coarse's netting more impressive (indicative of larger fish) than Mr. Powers's; Mr. Coarse's catch *clearly does* indicate $\theta > 12.2$ while Mr. Powers's catch *clearly does not*. These two fairly extreme cases of what is or is not indicated forms the basis for discriminating more generally which $\theta$ values have been indicated more or less well by "catches" with differently sized nets.

If one grants the plausibility of the principles that arise from looking at tests this way, then, according to the learning view I propose, one has the needed justification for distinguishing (where [BM] could not) an $\alpha$-significant rejection of $H$ with $T^+$-1600 from an $\alpha$-significant rejection of $H$ with $T^+$-25 (for any given $\alpha$). For, the magnitude of the discrepancy which the former (more powerful) tool indicates is smaller than the one the latter (coarser) test does.[12] The general principle for understanding what has been learned (i.e., the positively discrepant $\theta'$ value indicated) by a given (positive) observed difference $D_{\text{obs}}$ (between $\bar{X}_{\text{obs}}$ and a hypothesized value $\theta_H$) from a test $T^+$ is this:

(5.2) (i) $D_{\text{obs}}$ is a *good* indicator that $\theta$ exceeds $\theta'$ only if (and only to the extent that) $\theta'$ *infrequently gives rise to so large a difference*.
(ii) $D_{\text{obs}}$ is a *poor* indicator that $\theta$ exceeds $\theta'$ to the extent that $\theta'$ *frequently gives rise to such a large difference*.[13]

[12]Construing tests in terms of the magnitudes of the discrepancies detected shows the error in the common tendency to construe a statistically significant difference with a large sample size as *better evidence* against the null hypothesis than with a small sample size. That researchers have very often fallen prey to such a misinterpretation is, by now, well documented (e.g., Rosenthal and Gaito [1963] have demonstrated this in a group of psychological researchers). The misinterpretation stems from construing significance levels as *E-R* measures (of the plausibility of the null hypothesis). The smaller the significance level, the less plausible is $H$, and so the more plausible is its rejection; at least on such an *E-R* construal of significance levels. Coupling such a construal with the greater reliability accorded to experiments as the number of observations increases, explains the tendency to deem an $\alpha$-significant result with a large sample size as more impressive than one with a smaller sample size.

[13]It is by means of principle (5.2) that, on the learning view of testing I recommend, the results of orthodox tests should be interpreted. To get a more concrete look at what a specific interpretation might look like, fix the value of "infrequently" in (i) at .15. Suppose test $T^+$ is applied to the fish-length example and $D_{\text{obs}}$ is significant at level .02; i.e., $H$ is rejected with $T^+$ at level .02. We have the following:

**6. The Appropriateness of Orthodox Tests for Scientific Learning.** Even if it is agreed that orthodox-test conclusions can be reinterpreted in terms of the values of $\theta'$ discrepant from $\theta$ (for $T^+$, these are values in excess of $\theta_H$) that are or are not indicated, we still seem to obtain a negative answer to the question: Are tests which are "good" according to the criteria of low long-run error-rates also "good" from the point of view of the learning model? It would seem that a test is "good" for the purpose of learning about discrepancies (between actual and hypothesized values of $\theta$) if it classified an observed difference $D_{\text{obs}}$ statistically significant (and so grounds for rejecting $H$) just in case the difference was a "good indicator" (in the sense of (5.2(i))) of a *scientifically* (or substantively) important discrepancy. But, as the critics of orthodox tests have shown, it is possible to so bias a test against null hypothesis $H$ (i.e., make it so sensitive to discrepancies from $\theta_H$) that a rejection of $H$ from a test with error rates as low as one chooses still may be a *poor* indicator of a given scientifically important discrepancy. Rejecting $H$: $\theta = 12$, for example, with a .02-significant difference (with "best" test $T^+$) is a poor indicator of a discrepancy .2 inches from 12. And if .2 was deemed a negligibly small discrepancy, the test would fail miserably for the task of learning about non-negligible ones.[14]

---

(i) The values of $\theta'$ that give rise to a .02-significant difference (no more than) 15 percent of the time are those that exceed the hypothesized value, namely, 12 by (no more than) 1 $\sigma_{\bar{X}}$. That is, a .02-rejection of $H$ (with $T^+$) is a good indication that $\theta$ exceeds 12 + 1 $\sigma_{\bar{X}}$ (equivalently, that $\theta$ exceeds $\bar{X}_{\text{obs}} - 1\ \sigma_{\bar{X}}$.)

Focus now on .02-significant results from $T^+$-25 and from $T^+$-1600. Both test results are equally good indicators (using rule (i)) of a discrepancy in excess of 1 $\sigma_{\bar{X}}$ (equivalently, of $\theta$ in excess of 12 + 1 $\sigma_{\bar{X}}$.) Then, the result from $T^+$-25 is *as good an indication* that $\theta$ exceeds 12.4 inches, as is the result from $T^+$-1600 that $\theta$ exceeds 12.05 inches. Similarly, both results are *equally poor* indicators of positive discrepancies ($\theta - 12$) in excess of those that have an equally high frequency of giving rise to such significant differences. Letting .85 be used for "frequently" in (5.2) (ii), we have:

(ii) A .02-rejection of $H$ is a poor indicator that $\theta$ exceeds $\bar{X}_{\text{obs}} + 1\sigma_{\bar{X}}$. It follows that a .02-rejection of $H$: $\theta = 12$ with $T^+$-25 is *as poor* an indication that $\theta$ exceeds 13.2 as is a .02-rejection from $T^+$-1600 that $\theta$ exceeds 12.15.

Mathematically, (i) and (ii) are equivalent to: $\bar{X}$ is a good indicator that $\theta$ exceeds the value of the lower bound of an 85 percent confidence interval; and is a poor indicator that $\theta$ exceeds the value of the upper bound of an 85 percent confidence interval. The major difference is that under the learning model not all $\theta$ values in the corresponding upper and lower confidence intervals are on par. For example, far from construing $\theta$ values in excess of $\bar{X}_{\text{obs}}$, but less than the upper bound of a confidence interval, as acceptable estimates of $\theta$, such values are *poorly* indicated; the more poorly the further they exceed $\bar{X}_{\text{obs}}$. An analogous construal of "accept $H$" occurs in Mayo 1983.

[14]This is easy to see by referring to the cut-off points of (i) and (ii) in n. 13. That is, since we saw that a .02-significant result with $T^+$-1600 is a poor indicator of $\theta$ in excess of 12.15, it is an even poorer indicator that $\theta$ exceeds 12.2. Moreover, since a .02-significant result from $T^+$-25 is a good indication that $\theta$ exceeds 12.4, it is an even better indication that $\theta$ exceeds 12.2.

The reason that this negative conclusion appears inescapable is the tendency to equate the statistical conclusion with a substantive scientific one, and admittedly, orthodox tests are often formulated in a manner that encourages such an identification. What follows from such an equation is that a good statistical inference is equated with a good scientific one. But if a "good" scientific test is one that indicates all and only discrepancies of interest; then a test that is "good" from the point of view of low error-rates may fail to satisfy the criterion of scientific learning. This we have already seen. But by clearly distinguishing the statistical from the scientific conclusion, it is possible to critically interpret the former's bearing on the latter. In this way the criticism of overly sensitive tests points to a possible *misinterpretation* of the substantive or scientific import of a statistically significant result.

In other words, a "good test on the learning model" can be understood in two ways. To avoid ambiguity, two criteria must be distinguished:

(6.0) [LM]: (i) A *statistical testing procedure* is good iff one is able to objectively evaluate what has and has not been learned from a statistical conclusion (reject or accept $H$).[15]
[LM]: (ii) A *statistical test conclusion* (e.g., $T^+$ rejects $H$) is [poor] good for learning about a given discrepancy between $\theta$ and $\theta'$ to the extent that it is a [poor] *good indicator* that $\theta$ exceeds $\theta'$, (in the sense 49(5.2)).

The question "Is the orthodox theory of testing appropriate for objective scientific learning?" can now be expressed as: "Is the orthodox theory of testing a good procedure in the sense of [LM] (i)?" We can now reason to a positive answer to this question by connecting the separate results of this paper:

(6.1) (1) One can objectively evaluate what has and has not been learned from a statistical conclusion iff one can objectively ascertain what it does and does not indicate in the sense of (5.2).
(2) One can objectively evaluate what a test result does and does not indicate (in the sense of (5.2)) iff one can objectively determine the frequencies with which test results arise from various values of $\theta'$ (i.e., from various discrepancies between actual and hypothesized $\theta$ values.)

[15]An obvious additional requirement, of course, is that the procedure be capable of being used as a learning tool to begin with. A procedure that often informed one that little had been learned would otherwise count as "good" on [LM] (i). Orthodox tests, do, however, satisfy this additional requirement, since by suitably specifying tests they can be made to detect discrepancies of interest as frequently as one wants. More importantly, the learning construal of test results also indicates how to specify a subsequent test to learn more.

(Equivalently, a test result can be evaluated objectively iff one can objectively determine the error frequencies of a test result for any of the possible values of $\theta$ being considered.)

(3) An orthodox test (as defined in (2.7)) permits the objective evaluation of error frequencies of a test result over possible values of $\theta$.

(4) From (1)–(3) it follows that an orthodox test *procedure* (assuming its assumptions are approximately satisfied) is good for the task of objective scientific learning (as defined in (6.1) [LM] (i)); for it allows one to determine if particular results are good in the sense of [LM] (ii). And this is what I have sought to show.

**7. Conclusion and Suggestions for Further Work.** The view of testing that emerges in the learning model I propose is this: A test functions as a standard tool for detecting discrepancies between a hypothesized parameter or model, and the parameters or models *indicated* by the experimental observation (by means of such rules as (5.2)). Deliberate choices of error frequencies (or "operating characteristics") before the test serve to specify the type of discrepancies that the test is frequently able to detect.[16] In this respect the test functions much like an instrument for magnifying discrepancies typically indiscernible from a single sample.[17] Once the data are in hand, considerations of error properties (i.e., the test's power to detect various discrepancies) are far from being irrelevant (as critics often allege). The ability to guarantee or control error frequencies is what enables an orthodox test to function as a nonsubjective tool for understanding what observed test results indicate about their source.

[16]Giere (1976) suggests specifying tests so that they have appropriately high probabilities of detecting all and only discrepancies about which it would be of interest to learn in a given inquiry. By replacing the appeal to behavioristic considerations (of the "seriousness" of certain errors) with an appeal to scientific considerations of the magnitudes of discrepancies deemed important, Giere provides the groundwork for the reinterpretation of tests I suggest. According to Giere, considerations of scientifically important discrepancies can be based on the uses to which the test result is to be put; and as such, most investigators can agree on what counts as a scientifically important discrepancy.

On the present view, in contrast, whether or not specifications of scientifically important discrepancies are available, a "good" test procedure should enable one to understand what sort of discrepancies have in fact been detected by a given statistical result (i.e., it should satisfy [LM] (i)). Whether or not the resulting information (from the statistical test) is relevant for a given use is, on the present view, a separate question.

[17]Isaac Levi (1980), in construing Neyman-Pearson tests as "routinizable programs" and as using observations as "inputs" as opposed to as evidence for deliberation and inference, also views orthodox tests as instruments of a sort. My view of tests as learning tools may well fall under Levi's view of a "routine," if that notion is suitably broadened. (Admittedly, [LM] does not use data as evidence for deliberation if, as Levi maintains, this requires assigning posterior probabilities [or other E-R measures] to hypotheses.) In that case, I would see "routines" as the major (if not the only) methods needed for using data in the "enterprise of knowledge"—a view with which, I take it, Levi agrees.

In test $T^+$, for example, an observed difference is understood to indicate $\theta$ is positively discrepant from $\theta'$ by determining that almost none of the possible samples would give rise to such a large difference were $\theta$ no greater than $\theta'$. This understanding of a test result is neither a decision to behave as if $\theta$ exceeded $\theta'$, nor an assignment of evidential strength to the hypothesis that $\theta$ exceeds $\theta'$. Rather it is a statement of what has been learned about how differences (like the one observed) could be generated *systematically* (i.e., more often than accidental effects). The statement "By fishing in Lake B using $T^+$-25 (Mr. Coarse's net) a 'catch' (of 25 fish) with mean length as large as 12.8 inches would be generated systematically" asserts "The mean value of such catches in Lake B exceeds 12.2 inches."

This is a statement about the distribution of observations (i.e., $\bar{X}$ values) that *would arise* if the mean of all of the (25-fish) catches from Lake B exceeds 12.2 (i.e., iff the mean of $\bar{X}$ exceeds 12.2). For it is about this (sampling) distribution that the statistical hypotheses in (2.0) refers. Admittedly, the sequence of observations to which the test result refers is likely to be *hypothetical*. But this does not vitiate its use as a standard way of understanding the process generating the actual observation. By learning about the mean of the distribution $\bar{X}$, for example, one is at the same time learning about the initial scientific hypotheses ($\mathcal{H}$ and $\mathcal{J}$) as to whether $\theta$, the mean length of fish in a given population exceeds 12 inches by various amounts. The reason is that the mean of $\bar{X}$ is equal to the population mean $\theta$ (see note 3).

How does the learning model relate to more global statements of scientific knowledge? A full answer is beyond the scope of this paper, but my interpretation of test $T^+$ provides the beginning for the answer I would suggest. For the function served by statistical tests when the quantity $\theta$ represents mean fish lengths is much the same as when $\theta$ represents the length of a table, the mean concentration of a hormone, the mean increased fitnesses of a given species, the mean deflection of light near the sun, and so on. For whether it is due to the incompleteness and inaccuracies of observation and measurement or the variability of the effect or system of interest, experimental data is rarely expected to agree precisely with testable predictions; even when they are derived from scientific hypotheses that adequately describe the phenomenon of interest. As such, the testable prediction may be expressed as a statement about a distribution of observations that *would be expected;* that is, as a statistical hypothesis *about an experiment*. Statistical tests then serve to detect and distinguish observed differences that are due to accidental or trivial discrepancies, from those due to systematic or substantively important ones.

To arrive at most interesting statements of scientific knowledge several individual statistical tests have to be imbedded within a larger, more com-

plex model of a scientific learning-strategy. To this end, a system of "metastatistical" principles, along the lines of those developed in (5.2) for $T^+$, may be developed for a variety of statistical tests. Then, by means of principles spanning several different theories (e.g., theories of experiment, of observation, of the primary scientific phenomenon) individual tests may be specified, interpreted, crosschecked and corrected, by reference to other statistical tests within the larger model of the given learning-effort.

To the extent that the learning-model interpretation of tests that I suggest succeeds in capturing the appropriate function of statistical methods in science, orthodox tests avoid being dethroned by their critics. If I am correct (in thinking the extent is considerable), then the challenge would be for proponents of non-orthodox methods (e.g., Bayesians, fiducialists, etc.) to show the ability of their methods to accomplish the actual tasks of experimental learning described here. In any case, my challenge might at least encourage philosophers of statistics to weigh the merits and demerits of statistical methodologies by applying them to actual inquiries.

## REFERENCES


Birnbaum, A. (1977), "The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; With a Criticism of the Lindley-Savage Argument for Bayesian Theory", *Synthese 36*: 19–50.

Carnap, R. (1950), *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Edwards, W.; Lindman, H.; and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research", *Psychological Review 70*: 193–242.

Fetzer, J. H. (1981), *Scientific Knowledge*. Dordrecht: D. Reidel.

Fisher, R. A. (1955), "Statistical Methods and Scientific Induction", *Journal of the Royal Statistical Society B 17*: 69–78.

Giere, R. N. (1969), "Bayesian Statistics and Biased Procedures", *Synthese 20*: 371–87.

———. (1976), "Empirical Probability, Objective Statistical Methods and Scientific Inquiry", in *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science,* vol. 2, W. L. Harper and C. A. Hooker (eds.). Dordrecht: D. Reidel, pp. 63–101.

———. (1977), "Testing vs. Information Models of Statistical Inference", in *Logic, Laws and Life,* R. G. Colodny (ed.). Pittsburgh: University of Pittsburgh Press, pp. 19–70.

Good, I. J. (1950), *Probability and the Weighing of Evidence*. London: Griffin; New York: Hafner.

———. (1980), "The Diminishing Significance of a P-Value as the Sample Size Increases", *Journal of Statistical Computation and Simulation 11*: 307–9.

———. (1981), "Some Logic and History of Hypothesis Testing", in *Philosophy in Economics,* J. C. Pitt (ed.), Dordrecht: D. Reidel, pp. 149–74.

———. (1982), "Standardized Tail-Area Probabilities", *Journal of Statistical Computation and Simulation 13*: 65–66.

Hacking, I. (1965), *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

———. (1980), "The Theory of Probable Inference: Neyman, Peirce and Braithwaite", in *Science, Belief and Behavior: Essays in Honour of R. B. Braithwaite,* D. H. Mellor (ed.). Cambridge: Cambridge University Press, pp. 141–60.

Jeffreys, H. [1938] (1961), *Theory of Probability*. Oxford: Clarendon Press.

Kempthorne, O. (1971), "Probability, Statistics, and the Knowledge Business", in *Foundations of Statistical Inference,* V. P. Godambe and D. A. Sprott (eds.). Toronto: Holt, Rinehart and Winston of Canada, pp. 470–92.

Kempthorne, O., and Folks, L. (1971), *Probability, Statistics, and Data Analysis.* Ames: Iowa State University Press.

Kyburg, H. E., Jr. (1971), "Probability and Informative Inference", in *Foundations of Statistical Inference,* V. P. Godampe and D. A. Sprott (eds.). Toronto: Holt, Rinehart and Winston of Canada, pp. 82–103.

———. (1974), *The Logical Foundations of Statistical Inference.* Dordrecht: D. Reidel.

Levi, I. (1980), *The Enterprise of Knowledge.* Cambridge: The MIT Press.

Lindley, D. V. (1965), *Introduction to Probability and Statistics From a Bayesian Point of View. Part 2: Inference.* Cambridge: Cambridge University Press.

———. (1972), *Bayesian Statistics, A Review.* Philadelphia: Society for Industrial and Applied Mathematics.

Mayo, D. (1981a), "In Defense of the Neyman-Pearson Theory of Confidence Intervals", *Philosophy of Science 48*: 269–80.

———. (1981b), "Testing Statistical Testing", in *Philosophy of Economics,* J. C. Pitt (ed.). Dordrecht: D. Reidel, pp. 175–203.

———. (1982), "On After-Trial Criticisms of Neyman-Pearson Theory of Statistics", in *PSA 1982,* vol. 1, P. Asquith and T. Nickles (eds.). East Lansing: Philosophy of Science Association, pp. 145–58.

———. (1983), "An Objective Theory of Statistical Testing", *Synthese 57*: 297–340.

Neyman, J., and Pearson, E. S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypothesis", *Philosophical Transactions of the Royal Society A 231*: 289–337. (Reprinted in *Joint Statistical Papers,* Berkeley: University of California Press, 1967, pp. 276–83.)

Pearson, E. S. (1947), "The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 × 2 Table", *Biometrika 34*: 139–67. (Reprinted in *The Selected Papers of E. S. Pearson,* Berkeley: University of California Press, pp. 169–97.)

———. (1955), "Statistical Concepts in Their Relation to Reality", *Journal of the Royal Statistical Society B 17*: 204–7.

Rosenkrantz, R. D. (1977), *Inference, Method and Decision.* Dordrecht: D. Reidel.

Rosenthal, R., and Gaito, J. (1963), "The Interpretation of Levels of Significance by Psychological Researchers", *Journal of Psychology 55*: 33–38.

Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference.* Dordrecht: D. Reidel.

Smith, C. (1977), "The Analogy between Decision and Inference", *Synthese 36*: 71–85.

Spielman, S. (1973), "A Refutation of the Neyman-Pearson Theory of Testing", *British Journal for the Philosophy of Science 24*: 201–22.