DEBORAH G. MAYO

DID PEARSON REJECT THE NEYMAN–PEARSON PHILOSOPHY OF STATISTICS?*

ABSTRACT. I document some of the main evidence showing that E. S. Pearson rejected the key features of the behavioral-decision philosophy that became associated with the Neyman–Pearson Theory of statistics (NPT). I argue that NPT principles arose not out of *behavioral* aims, where the concern is solely with behaving correctly sufficiently often in some long run, but out of the *epistemological* aim of learning about causes of experimental results (e.g., distinguishing genuine from spurious effects). The view Pearson did hold gives a deeper understanding of NPT tests than their typical formulation as 'accept-reject routines', against which criticisms of NPT are really directed. The 'Pearsonian' view that emerges suggests how NPT tests may avoid these criticisms while still retaining what is central to these methods: the control of error probabilities.

1. INTRODUCTION

The Neyman–Pearson Theory of statistics (NPT), often referred to as 'standard' or 'orthodox' statistical theory, is the generally-received view in university departments of statistics, and it underlies common statistical reports. Strictly speaking, NPT procedures of hypotheses testing and estimation are only a part of the full collection of methods referred to as 'sampling theory', which also includes methods of experimental design and data analysis. But it is this part on which philosophical critics of 'standard' or 'orthodox' statistical theory have generally concentrated. Egon S. Pearson (not to be confused with his father, Karl[1]), although one of the two founders of NPT, rejected the statistical philosophy that ultimately became associated with NPT, or so I shall argue. Because specific citations are important for my case, I shall quote throughout at some length. Another reason for doing so is to put these remarks – largely overlooked in discussions of the philosophy of statistics – together in one place.

Understanding Pearson's rejection of the NPT philosophy is of more than merely historical interest. It is also highly relevant to the allegations of many philosophers of statistics – Fetzer (1981), Hacking (1965) (but compare Hacking (1980)), Howson and Urbach (1989), Kyburg (1971, 1974), Levi (1980), Rosenkrantz (1977), Seidenfeld (1979), Spielman (1973), and of several statisticians as well – that NPT,

Synthese **90**: 233–262, 1992.

despite its widespread use, is inappropriate for statistical inference in science. In statistical practice as well, there continues to be a lively debate over the use of NPT methods, with their seeming rigidities, in the face of the vicissitudes of actual experimental data (e.g., in clinical trials and risk assessments). Many of these contemporary criticisms mirror, I claim, Pearson's own reasons for rejecting the philosophy typically associated with NPT. Extricating the view Pearson *did* hold, I think, gives a much deeper understanding of NPT principles than that found in statistics texts, against which criticisms of NPT are really directed. Such an understanding suggests how NPT may avoid some of these criticisms while still retaining what is central to sampling theory methods: the fundamental importance of error probabilities. Finally, the 'Pearsonian' view of statistical inference that emerges seems to offer a promising avenue for using statistical reasoning to accomplish the task at which 'inductive logics' fell short: illuminating the nature and rationale of experimental learning in science.

2. NEYMAN–PEARSON THEORY OF STATISTICAL TESTS (NPT TESTS)

2.1. *Basic Notions*

I focus here on NPT tests. The mathematics of this testing theory defines functions on *experiments* modelled by statistical variables. The functions map possible values of these variables (i.e., possible experimental outcomes) to various hypotheses about the population from which outcomes may have originated. Commonly, the hypotheses are assertions about some property of this population, a *parameter*, which governs the statistical distribution of the experimental variable. For example, the statistical variable in a coin-tossing experiment might be the proportion of heads in n tosses, and the hypotheses, assertions about the (binomial) parameter p, the probability of heads on each toss. The NPT test splits the possible parameter values into two: one representing the *test hypothesis* H, the other the set of *alternative hypotheses* J. H, for example, might assert that $p = 0.5$, while J, that $p > 0.5$. (H here is *simple*, while J is *composite*.) The test maps the possible outcomes – the *sample space* – into either H or J; those mapped into H (i.e., into 'accepting' H) form the *acceptance region*, while those mapped into alternative J, the *rejection* (*of* H) *region*. This partition of the sample space is typically performed by specifying a cutoff point or *critical boundary*, beyond which an outcome enters the rejection region. An example would be to reject H whenever the sample proportion of heads is at least 0.8. Leaving these acceptances and rejections uninterpreted, the formalism of the NPT model simply describes the partitioning that results from the mapping rules as illustrated below:

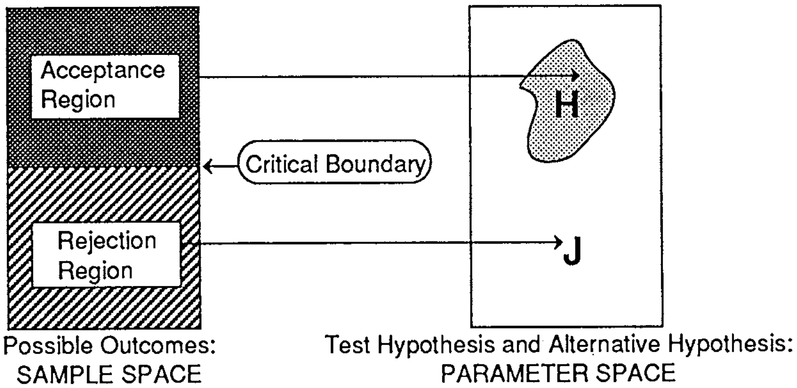
Acceptance Region
Critical Boundary
H
Rejection Region
J
Possible Outcomes: SAMPLE SPACE
Test Hypothesis and Alternative Hypothesis: PARAMETER SPACE

Fig. 1. NPT Tests as Mapping Rules.

The focus of the NPT test is on the probabilistic properties of these mapping rules, that is, on the probabilities that the rule would map to one or another hypothesis, under varying assumptions about the true parameter value. Two types of errors are considered: first, the test leads to reject H (accept J) even though H is true (the *Type I error*); and second, the test leads to accept H although H is false (the *Type II error*). The test is specified so that the probability of a Type I error, represented by α, may be fixed at some small number, such as 0.05, or 0.01. In other words, the test is specified so as to ensure it is very improbable for a certain result to occur; namely, an outcome falls in the 'rejection (of H) region' although the hypothesis H is correct. Having fixed α, called the *size* of the test, NPT principles seek out the test which at the same time has a small probability, represented by β, of committing a Type II error: accepting H, when J is actually the correct hypothesis. $1 - \beta$ is the corresponding *power* of the test. (When alternative J contains more than a single value of the parameter, i.e., when J is *composite*, the value of β varies according to which alternative in J is true.) α and β are the test's *error probabilities*; they are not probabilities of hypotheses, but the probabilities (in a frequentist sense) with which certain results would occur in a long-run sequence of applications of such test rules.

This leads to the cornerstone of NPT tests: their ability to ensure that a test's error probabilities will not exceed some suitably small values, fixed ahead of time by the user of the test, regardless of which hypothesis is correct. These key points can be summarized as follows:

> An *NPT test* (of hypothesis H against alternative J) 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 hypothesis) the rule will erroneously reject H and erroneously accept H no more than α(100%) and β(100%) of the time, respectively.

The 'best' test of a given size α (if it exists) is the one that at the same time minimizes the value of β (equivalently, maximizes the power) for all possible alternatives J.

2.2. *Behavioral Decision Philosophy of NPT: Tests as Accept-Reject Routines*

The proof by Neyman and Pearson of the existence of 'best' tests encouraged the view that tests (particularly 'best' tests) provide the scientist with a kind of *automatic rule* for testing hypotheses. Here tests are formulated as mechanical rules or 'recipes' for reaching one of two possible decisions: 'accept hypothesis H' or 'reject H (accept alternative J)'. The justification for using such a rule is its guarantee of specifiably low error rates in some long run.

This interpretation of the function and the rationale of tests was well suited to Neyman's statistical philosophy. For Neyman, "[t]he problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action: either take step A or take step B," (Neyman 1950, p. 258). These are not decisions to accept or to believe that what is hypothesized is (or is not) true, Neyman stresses; rather, "to accept a hypothesis H means only to *decide to take action A rather than action B*" (ibid., p. 259; emphasis added). On Neyman's view, when evidence is inconclusive all talk of 'inferences' and 'reaching conclusions' should be abandoned. Instead, Neyman sees the task of a theory of statistics as providing rules to guide our behavior so that we shall avoid making erroneous decisions too often in the long run of experience. A clear statement of such a rule is the following:

> 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 $x \leqslant x_0$, accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leqslant x_0$ 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)

Tests interpreted as such rules of *inductive behavior* yield the *behavioristic model* of tests, typically associated with Neyman and Pearson. The question is this: Are tests that are 'good' according to the behavioristic criteria (of low error-probabilities) also good for obtaining scientific knowledge? That is, are they good for finding out *what is the case*, as opposed to *how it is best to behave?* Most philosophers of statistics say no.

It is admitted that the orthodox test may be sensible, if one is in the sort of decision-theoretic context envisioned by the behavioristic approach. The paradigm example is acceptance sampling in industrial quality control. Here the choice is whether or not to reject a certain batch of products as containing too many defectives, say, for shipping. This is a paradigmatic case in which the primary interest is ensuring that the long-run risks of such business decisions are no more than can be 'afforded', and in such cases, NPT can provide the desired guarantees. But testing claims in scientific contexts does not seem to be like this. As Henry Kyburg aptly put it:

> To talk about accepting or rejecting hypotheses . . . is *prima facie* to talk epistemologically; and yet in statistical literature to accept the hypothesis that the parameter μ is less than μ^* is often merely a fancy and roundabout way of saying that Mr. Doe should offer no more than $36.52 for a certain bag of bolts . . . . (Kyburg 1971, pp. 82–83)

This is true about the behavioral model of NPT, in which a test result is interpreted as taking an action, e.g., paying a certain price for bolts. But this is not, I claim, the only, nor even the intended, interpretation of NPT test results.

2.3. *An Evidential Interpretation of NPT: Birnbaum's Confidence Concept*

Alan Birnbaum (1969, 1977) had argued that NPT admits of two types of interpretations: on one, Neyman's behavioral decision view, the test result is literally a decision *to act* in a certain way; on the other, which Birnbaum called an "evidential" view, the test result is interpreted as providing strong or weak evidential support for one or another hypothesis. He called the concept underlying this evidential interpretation of NPT the *confidence concept* which he formulated as follows:

> (Conf): A concept of statistical evidence is not plausible unless it finds 'strong evidence for J as against H' with small probability (α) when H is true, and with much larger probability ($1 - \beta$) when J is true.[2] (Birnbaum 1977, p. 24)

Birnbaum argued that scientific applications of NPT made intuitive use of something like the confidence concept and, although he felt that such concepts have not been incorporated explicitly in NPT (or any other statistical theory), he found clues of these non-behavioral intuitions in the writings of Pearson. One interesting document Birnbaum (1977, p. 33) supplies is an unpublished note by Pearson, commenting in 1974 on an earlier draft of Birnbaum's own paper:

> 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. (emphasis added)

Birnbaum, I believe, was correct to identify in Pearson a tendency to view the behavioral model of NPT as a heuristic device, serving to communicate what the tests could be used for, but requiring reinterpretation in scientific contexts. However, I do not think that Birnbaum's system, so far as he worked it out,[3] in which NPT results are reinterpreted in terms of strong or weak evidence for hypotheses, captures Pearson's divergence from the NPT philosophy.

2.4. *NPT Philosophy: the Function and Rationale of Tests*

As NPT formally developed in a decision-theoretic framework (along with the work of Wald), the NPT statistical philosophy has generally been taken as the behavioral decision one (Section 2.2). I want now to examine two closely-connected aspects of this decision philosophy: first, the justification of tests in terms of low (long-run) error rates and, second, the function of tests as routine decision rules. Both are at the heart of epistemological criticisms of NPT; they seem to lead to Neyman's view that a test "does not contribute anything about the falsehood or correctness of" hypotheses.[4]

(i) *Long-Run* (*Low Error-Probability*) *Justification*: Since the criteria for goodness of a test are its low error probabilities in the frequentist sense, the justification for using tests is solely in terms of their ability to guarantee low long-run errors in some sequence of applications. This is not a final measure of probability of hypotheses. To reject H, for example, with a test having a low probability of erroneous rejections does not say the specific rejection has a low probability of being in error, but only that it arises from a testing procedure which has a low probability of leading to erroneous rejections. So, what is the rationale, it may be asked, for deeming a specific rejection of H as counter indicating hypothesis H?

(ii) *Tests as Decision 'Routines' with Pre-specified Error Properties*: The NPT decision model does not give an interpretation customized to the specific result realized: a result either is or is not in the pre-specified rejection region. But, intuitively, if a given test rejects H with an outcome several standard deviations beyond the critical boundary (between rejection and acceptance of H), there is an indication of a greater discrepancy from H than if the same test rejects H with an outcome just at the critical boundary. Both, however, are identically reported as reject H (and accept some alternative J), and the probability of a Type I error (the test's pre-specified size) is identical for any such rejection.[5] On this model, as Isaac Levi puts it, NPT tests are means "for using observation reports as inputs into programs designed to select acts" (Levi 1980, p. 406) as opposed to using them as *evidence* in deliberation.

These features, taken as integral to a strict reading of the NPT model, underlie contemporary criticisms of NPT, as well as much of the original attack by R. A. Fisher. In his grand polemic style, Fisher declared that followers of the behavioristic approach are like

> Russians (who) are made familiar with the ideal that research in pure science can and

> should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation. (Fisher 1955, p. 70)

A similar comparison is made with the United States:

> In the U.S. also the great importance of organized technology has I think made it *easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money.* (Ibid., p. 70)

The allegation is essentially the one cited earlier (e.g., by Kyburg): NPT methods seem suitable for industrial acceptance sampling, but not for drawing inferences in science. (Much more needs to be said to explain and respond to contemporary criticisms of NPT, something attempted elsewhere, e.g. in Mayo (1982, 1983, 1985, 1988).) But contemporary critics seem to have overlooked Pearson's deliberate response to Fisher's attacks. Perhaps this is because it occurs in an obscure, very short (but fascinating) paper, 'Statistical Concepts in their Relation to Reality' (Pearson 1955), not found in *The Selected Papers of E. S. Pearson.*

3. PEARSON REJECTS THE NEYMAN–PEARSON PHILOSOPHY

3.1. *Pearson's Heresy*

What one discovers in Pearson's (1955) response to Fisher (and elsewhere in his work) is that for scientific contexts Pearson rejects both the low long-run error probability rationale, and the non-deliberational, routine use of tests. These two features are regarded as so integral to the NPT model that, along with Birnbaum and other philosophers of statistics, let us grant they are primary components of the strict Neyman–Pearson philosophy. But, then, I think it is fair to say that Pearson himself rejected the Neyman–Pearson philosophy (but not NPT methods). Pearson did not publish much on his own statistical philosophy per se, but evidence scattered throughout his statistical papers offers a fairly clear picture of the rationale underlying his rejection of these decision features of NPT.

Let us begin with Pearson's (1955) response to Fisher's criticism. He insists that

> [t]here was no sudden descent upon British soil of Russian ideas regarding the function of science in relation to technology and to five-year plans. It was really much simpler – or worse. The original heresy, as we shall see, was a Pearson one! (Pearson 1955, p. 204)

Interestingly, Fisher directs his attacks at Neyman's behavioral approach, leaving Pearson out of it.[6] Nevertheless, Pearson protests here that the "original heresy" was his (i.e., "was a Pearson one")! Pearson does *not* mean it was he who endorsed the behavioral-decision model that Fisher attacks. The "original heresy" refers to the break Pearson made (with Fisher) in insisting tests explicitly take into account alternative hypotheses, in contrast to Fisherian significance tests, which did not. With just the single hypothesis (the null hypothesis) of Fisherian tests, there were many ways to specify the test, rendering the choice too arbitrary. With the inclusion of a set of admissible alternatives to H, it was possible to consider Type II as well as Type I errors, and thereby to constrain the appropriate tests.

So the central thing to see about Pearson's response to Fisher is that Pearson was not merely arguing that NPT methods can be interpreted in a manner other than a pragmatic behavioral-decision one, he was claiming that their original formulation (admittedly 'heretical' in the above sense) was not at all intended to capture decision-theoretic aims, aims which came later.

> Indeed, to dispel the picture of the Russian technological bogey, I might recall how certain early ideas came into my head as I sat on a gate overlooking an experimental blackcurrant plot . . .! (Ibid., p. 204)

To this marvelous depiction of Pearson sitting on a gate, Pearson adds a description of his earnest intent:

> *To the best of my ability I was searching for a way of expressing in mathematical terms what appeared to me to be the requirements of the scientist in applying statistical tests to his data.* After contact was made with Neyman in 1926, the development of a joint mathematical theory proceeded much more surely; *it was not till after the main lines of this theory had taken shape* with its necessary formalization in terms of critical regions, the class of admissible hypotheses, the two sources of error, the power function, etc., *that the fact that there was a remarkable parallelism of ideas in the field of acceptance sampling became apparent. Abraham Wald's contributions to decision theory of ten to fifteen years later were perhaps strongly influenced by acceptance sampling problems, but that is another story.* (Ibid., pp. 204–05; emphasis added)

Pearson proceeds to 'Fisher's next objection': to the terms "acceptance" and "rejection" of hypotheses, and to the Type I and Type II errors. His admission is revealing of his philosophy:

> It may be readily agreed that in the first Neyman and Pearson paper of 1928, more space might have been given to discussing how the scientific worker's attitude of mind could be related to the formal structure of the mathematical probability theory .... *Nevertheless it should be clear from the first paragraph of this paper that we were not speaking of the final acceptance or rejection of a scientific hypothesis on the basis of statistical analysis* .... Indeed, from the start we shared Professor Fisher's view that in scientific enquiry, *a statistical test is 'a means of learning'* .... (Ibid., p. 206; emphasis added)

So for Pearson the NPT framework, with its consideration of alternative hypotheses, was an outgrowth of an attempt to provide the tests then in use with an epistemological rationale, one based on their function as learning tools. Pearson clearly distances the mathematical apparatus from the later behavioral-decision construal to which Fisher objected, declaring in the final line of this paper that

> Professor Fisher's final criticism concerns the use of the term 'inductive behavior'; this is Professor Neyman's field rather than mine. (Ibid., p. 207)

3.2. *Pearson Rejects the Long-run Rationale*

It seems clear that for Pearson, the value of NPT tests (in scientific or learning contexts) need *not* lie in the long-run error-rate rationale found in the decision model. Pearson raises the question as follows, with a mention of 'inference' already in contrast with Neyman:

> How far then, can one go in giving precision to a philosophy of statistical inference? ... (Pearson 1947, p. 172)

He considers the rationale that might be given to NPT tests in two types of cases, A and B:

> (A) At one extreme we have the case where repeated decisions must be made on results obtained from some routine procedure ....
> (B) At the other is the situation where statistical tools are applied to an isolated investigation of considerable importance .... (Ibid., p. 170)

In cases of type A, long-run results are clearly of interest, while in cases of type B, repetition is impossible or irrelevant. For Pearson's treatment of the latter case (type B) the following passage is telling:

> 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 judgment*?

> 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 judgment which we control at a low figure? (Ibid., p. 173; emphasis added)

Regrettably, Pearson leaves this tantalizing question unanswered, claiming, "On this I should not care to dogmatize". Nonetheless, in studying how Pearson treats cases of type B, it becomes evident that in his view, "the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgment". In addressing this issue, Pearson intends to preempt the ('commonsense') criticism of long-run justifications of precisely the sort lodged by contemporary critics of NPT:

> Whereas when tackling problem A it is easy to convince the practical man of the value of a probability construct related to frequency of occurrence, in problem B the argument that 'if we were to repeatedly do so and so, such and such result would follow in the long run' *is at once met by the commonsense answer that we never should carry out a precisely similar trial again.*
>
> Nevertheless, it is clear that the scientist with a knowledge of statistical method behind him can make his contribution to a round-table discussion .... (Ibid., p. 171)

In seeing how, we are at once led toward substantiating my second claim that Pearson rejects the routine use and interpretation of NPT tests found in the behavioral model. For the scientist's contribution requires using tests to learn about causes – something which cannot be reduced to routines.

3.3. *Pearson On Non-routine Uses of Tests: An Example of Type B*

The notion that a primary function of statistical tests is their ability to teach us about causes by answering a series of standard questions, found throughout Pearson's work, is summarized in the opening of a 1933 paper, jointly written with Wilks:

> Statistical theory which is not purely descriptive is largely concerned with the development of tools which will assist in the determination from observed events of the probable nature of the underlying cause system that controls them .... We may trace the development through a chain of questionings: Is it likely, (a) that this sample has been drawn from a specified population, P; (b) that these two samples have come from a common but unspecified population; (c) that these k samples have come from a common but unspecified population? (Pearson and Wilks 1933, p. 81)

Consider the following example Pearson gives of a case of type B, where no repetition is intended:[7]

> *Example of type B.* Two types of heavy armour-piercing naval shell of the same calibre are under consideration; they may be of different design or made by different firms . . . . Twelve shells of one kind and eight of the other have been fired; two of the former and five of the latter failed to perforate the plate . . . . (Pearson 1947, p. 171)

The variable observed (i.e., the statistic) is the difference, D, between the proportions that perforate the plate from the two types of shell. Its observed value, D_{obs}, equals 11/24 (i.e., 10/12 – 3/8). Tests aid the scientist's "contribution to a round-table discussion", Pearson suggests, by informing of the result's cause, that is, by answering a question under (b), about the origin of the two samples of naval shells:

> Starting from the basis that individual shells will never be identical in armour-piercing qualities, however good the control of production, he has to consider how much of the difference between (i) two failures out of twelve and (ii) five failures out of eight is likely to be due to this inevitable variability. (Ibid., p. 171)

Notably, Pearson does not simply report whether or not this observed difference falls in the rejection region (i.e., whether a test maps it to 'reject H'), but calculates the probability "of getting as great or a greater positive difference" (ibid., p. 192) if hypothesis H were true – if there was no difference in piercing qualities. This is the *significance level* (Fisher's *p-level*) of the observed difference – a measure that clearly depends on the actual result observed.

The causal function of tests that Pearson intends leads to what is perhaps the strongest evidence to substantiate my claim that Pearson rejects the core of the NPT decision model: in striking contrast to the decision model, Pearson suggests that little turns on which of the different tests available one chooses to employ. Treating the (difference between two proportions) case in one way,[8] Pearson obtains an observed significance level of 0.052; treating it differently (along Barnard's lines), he gets 0.025 as the (upper) significance level. Pearson suggests that in important cases, the difference in error probabilities, depending on which of these tests is chosen, makes no real difference to substantive judgments in interpreting the results. It would make a difference, says Pearson, only in an automatic, routine use of tests:

> Were the action taken to be decided automatically by the side of the 5% level on which the observation point fell, it is clear that the method of analysis used would here be of

> vital importance. *But no responsible statistician, faced with an investigation of this character, would follow an automatic probability rule.* (Ibid., p. 192; emphasis added)

This is important because it enables Pearson to avoid the often raised criticism that since different choices of a test's error probabilities may yield different – even opposite – hypothesis appraisals, and since selecting the error probabilities is somewhat arbitrary, there is a great deal of arbitrariness in the results. For on Pearson's view, this would yield no inconsistency, as long as one correctly understands the different meanings that should be attached to the results of different tests. Each is effectively asking a different question. With respect to the two approaches considered here, Pearson goes on to say that[9]

> [t]he result of either approach would raise considerable doubts as to whether the performance of the first type of shell was as good as that of the second . . . . (Ibid., p. 192)

Surprisingly, the same type of admonishment against an 'automatic' use of tests, along with other remarks redolent of Pearson's 'inferential' use of tests, occur not just in Pearson's own papers, but in one or two of the joint papers of Neyman and Pearson. In 1928, for example, 'they' wrote:

> If then a statistician thoughtlessly decides, whatever be the test, to reject an hypothesis when $P \leqslant .01$, say, and accept it when $P > .01$, it will make a considerable difference to his conclusions whether he uses [one test statistic or another]. But as the ultimate value of statistical judgment depends upon a clear understanding of the meaning of the statistical tests applied, the difference between the values of the two P's should present no difficulty. (Neyman and Pearson 1928, p. 18)

(P here is equal to the significance level.) In other words, if the decision model of NPT is taken literally, one accepts or rejects H according to whether or not the observed outcome falls in the preselected rejection region. Just missing the cutoff for rejection, say, because the observed significance level is 0.06 while the fixed level for rejection is 0.05, automatically makes the difference between an acceptance and a rejection of H. The 'Pearsonian' view rejects such automation in scientific contexts because

> it is doubtful whether the knowledge that [the observed significance level] was really 0.03 (or .06) rather than .05 . . . would in fact ever modify our judgment when balancing the probabilities regarding the origin of a single sample. (Ibid, p. 27)

Most significant in this joint contribution is the declaration that

> [i]f properly interpreted we should not describe one [test] as more *accurate* than another, but according to the problem in hand should recommend this one or that as providing information which is more *relevant* to the purpose. (Ibid., pp. 56–7)

This introduces a criterion above and beyond low error rates, namely, the 'relevance' of the information. In addition, clues emerge for connecting tests (used nonroutinely) to what clearly sounds like inferences about causes:

> [T]he tests should only be regarded as tools which must be used with discretion and understanding.... we must not discard the original hypothesis until we have examined the alternative suggested, and have satisfied ourselves that it does involve a change in the real underlying factors in which we are interested... that the alternative hypothesis is not error in observation, error in record, variation due to some outside factor that it was believed had been controlled, or to any one of many causes.... (Ibid., p. 58)

The very title of the joint paper in which these remarks are made – 'On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference' – is itself clearly at odds with Neyman's decision philosophy. Puzzlement at this paper's distinct 'Pearsonian' flavor is removed if one spots a small and highly interesting note by Neyman at the end of this paper. It seems the 'joint' paper was largely a contribution by Pearson!

> I feel it necessary to make a brief comment on the authorship of this paper. Its origin was a matter of close co-operation, both personal and by letter.... Later I was much occupied with other work, and therefore unable to co-operate. The experimental work, the calculation of tables and the developments of the theory of Chapters III and IV are due solely to Dr. Egon S. Pearson. (Neyman and Pearson 1928, p. 66; signed by J. Neyman)

4. PEARSONIAN PHILOSOPHY OF EXPERIMENTAL LEARNING

4.1. *Three Steps in the Original Construction of NPT Tests*

Pearson's discussion of the steps involved in the original construction of NPT tests brings out key differences between Pearson's and Neyman's philosophies and, at the same time, allows one to pinpoint the key difference between the NPT 'sampling' framework and non-sampling approaches. After setting up the test (or null) hypothesis, and the alternative hypotheses against which "we wish the test to have maximum discriminating power" (Pearson 1947, p. 173), Pearson defines three steps in test constructions:

> *Step 1.* We must specify [the *sample space*,[10]] the set of results which could follow on repeated application of the random process used in the collection of the data . . . .
>
> *Step 2.* We then divide this set [of possible results] by a system of ordered boundaries . . . such that as we pass across one boundary and proceed to the next, we come to a class of results which makes us more and more inclined, the information available, to reject the hypothesis tested in favour of alternatives which differ from it by increasing amounts. (Ibid., p. 173)

Results make us "more and more inclined" to reject H as they get further away from the results expected under H; that is, as the results become more probable under the assumption that some alternative J is true than under the assumption that H is true. The probability (or density) of a result e given H is called the *likelihood of H* given e. We are 'more inclined' toward J as against H to the extent that J is more likely than H given e.

NPT requires a third step – to ascertain the error probability associated with each measure of disinclination (each 'contour level'):

> *Step 3.* We then, if possible, associate with each contour level the chance that, if [H] is true, a result will occur in random sampling lying beyond that level.[11] (Ibid.)

For example, Step 2 might give us the likelihood or the ratio of likelihoods of hypotheses given evidence, i.e., the likelihood ratio. At Step 3 the likelihood ratio is itself treated as a statistic, a function of the data with a probability distribution. This enables calculating, for instance, the probability of getting a high likelihood ratio in favor of H, as against a specific alternative J′, when in fact the alternative J′ is true, i.e., an error probability. Learning that this probability is high counts against taking high likelihood for H as indicating the truth of H (as against J′). An analogy can be made with an examination score: Step 2 gives the score; Step 3 considers how frequently such a score would arise under various hypotheses, say, about what proportion of some material the person tested knows. Step 3 might tell us, for example, that a score of 65% ('passing') would very frequently occur on the test in question even if the subject knew only 20% of the material. As with a statistical test, this counts against taking a passing score as a good indication that a subject knows most of the material. (This analogy, we shall see, also suggests the manner in which error probabilities at Step 3 function in learning about causal origins.)

Pearson explains that in the original test model Step 2 (using likelihood ratios) did precede Step 3, and that only afterward did the NPT model start with the fixed error value for Step 3 and then determine the associated contour (i.e., the critical bounds for the rejection region). Pearson warns that:

> although the mathematical procedure may put Step 3 before 2, we cannot put this into operation before we have decided, under Step 2, on the guiding principle to be used in choosing the contour system. That is why I have numbered the steps in this order. (Ibid., p. 173)

However, if the rationale is *solely* long-run error probabilities, one loses sight of Step 2. That is why it is invisible in the standard decision construal of NPT. On this construal, having set up the hypotheses and sample space (Step 1), there is a jump to Step 3, fixing the error probabilities, on the basis of which a good (or best) NPT test indicates which outcomes to map into 'reject H' (the rejection region). In a sense the result of Step 3 automatically accomplishes Step 2: it describes how the test, selected for its error probabilities, is dividing the possible outcomes. But this is different from having first deliberated at Step 2 as to which outcomes are 'further from' or 'closer to' H in some sense, and thereby *should* incline us more or less to reject H. The resulting test, while having low error probabilities, may fail to ensure that the test has an increasing chance of rejecting H the more the actual situation deviates from the one H hypothesizes. Many counterintuitive NPT tests arise, e.g., certain mixed tests,[12] I believe, because tests are couched in the decision framework in which the task Pearson intended for Step 2 is absent.[13]

4.2. *Likelihood Principle vs. Error Probability Principles*

However, it might be asked, if Pearson is so concerned with Step 2, *why* go on to include Step 3 in the testing model at all? In other words, if Pearson is interested in how much a result 'inclines us' to reject H, *why* not just stop after providing a measure of such inclination at Step 2, instead of going on to consider error probabilities at Step 3? Indeed, this is precisely what many critics of NPT have asked. This was essentially Ian Hacking's (1965) point about NPT. Hacking argued that the *likelihood ratio* (of H against alternative J) provides an appropriate measure of *support* for H against J (a view he later came to doubt[14]). On such a likelihood view (he called it the Law of Likelihood) the tests *should* just report the measure of support or inclination (at Step 2) given the data. This conveys the full impact of the data, so there is no need to go on to consider the probability distribution of the support measure itself (at Step 3). This last probability, being assigned over the sample space, is an example of a *sampling distribution* (in this case, of the likelihood ratio statistic), which is why NPT is called a 'sampling theory'.[15]

This points out the crucial difference between NPT and non-sampling approaches, such as Bayesian or likelihood accounts: while in the former approach the (sampling) distribution is viewed as having crucial importance in interpreting results, e.g., in Step 3, in the latter, the relevant evidence contributed by the data is fully contained in the likelihood ratio actually obtained – a point formally expressed in the *likelihood principle.* According to the likelihood principle, which underlies Bayesian and likelihood accounts,[16] as D. V. Lindley remarks:

> if we have 2 pieces of data . . . with the same likelihood function . . . the inferences about [μ] from the two data sets should be the same. This is not usually true in the orthodox [NPT] theory, and its falsity in that theory is an example of its incoherence . . . . As Jeffreys has said, what has what might have happened, but did not, got to do with inferences from the experiment? (Lindley 1976, p. 361)

From the Bayesian point of view, the interest in what "might have" occurred renders NPT 'incoherent'; but, from the NPT point of view, Bayesian (and other non-sampling) methods are unpalatable just because they *ignore* what the data generation procedure might have produced. NPT error probabilities can only be derived from sampling distributions, the very distributions a committed Bayesian is, of course, happy to ignore as irrelevant for inference:[17]

> It is methods that are not based on the likelihood function that are suspect. In particular, unbiased estimates, minimum variance properties, sampling distributions, significance levels, power, *all depend on something more – something that is irrelevant in Bayesian inference – namely the sample space.* (Lindley 1971, p. 436; emphasis added)

A number of contemporary criticisms of NPT concepts echo the same theme. Strictly speaking, such 'criticisms' are really expressions of the incompatibility of two aims: that of providing methods (for testing and estimation) with specifiable error probabilities, and that of providing a measure ('after the trial'), e.g., the likelihood ratio, posterior probabilities to quantify support, belief, etcetera. Simply noting their inconsistency leaves unanswered the question of whether one is more appropriate for a given task. The appropriateness of NPT for science turns on showing why control of error probabilities is so important to experimental learning. Here is where Pearson's rejection of the long-run rationale of error probabilities and his nonroutine use of tests (from Section 3) come together with the Pearsonian logic of test construction from Section 4.1.

4.3. *Likelihoods Alone (Step 2) are Insufficient for Pearsonian Reasoning*

Pearson's explanation of why he and Neyman deemed the error probability calculations of Step 3 so essential is *not* a pragmatic decision concern with low error rates (in the long run of business), but a concern with learning from experiments. Reflecting on this question (in 'Some Thoughts on Statistical Inference'), Pearson (1962, p. 277) tells of their "dissatisfaction with the logical basis – or lack of it – which seemed to underlie the choice and construction of statistical tests", explaining that he and Neyman "were seeking how to bring probability theory into gear with the way we think as rational human beings":

> But looking back I think it is clear why we regarded the integral of probability density within (or beyond) a contour as more meaningful than the likelihood ratio – more readily brought into gear with the particular process of reasoning we followed.
>
> *The reason was this. We were regarding the ideal statistical procedure as one in which preliminary planning and subsequent interpretation were closely linked together – formed part of a single whole.* It was in this connexion that integrals over regions of the sample space were required. *Certainly, we were much less interested in dealing with situations where the data are thrown at the statistician and he is asked to draw a conclusion. I have the impression that there is here a point which is often overlooked* . . . . (Ibid., pp. 277–78; emphasis added)

I have the impression that Pearson is correct. Perhaps because the philosophical problem for a theory of statistics is typically posed as how given data relate to hypotheses, (and because texts present statistical methods separately from those of experimental design), the main focus of philosophical discussions is on what rival approaches tell one to do once "data are thrown at the statistician and he is asked to draw a conclusion"; e.g., accept or reject for a NPT test or compute a posterior probability for a Bayesian.

Why do error probabilities become relevant when the focus turns to the 'preliminary planning'?[18] I suggest the reasons are these. For one, by considering ahead of time the chances a given experiment has of detecting discrepancies of interest, one can avoid carrying out a study with little or no chance of teaching one what one wants to learn; for example, one can determine ahead of time how large a sample would need to be in a certain test to have a reasonably high chance (power) of rejecting H when in fact some alternative J is true. I take it that few would dispute this (before-trial) function of error probabilities. But error probabilities are also relevant for interpreting the particular results after the trial, and it is on this claim that I want to focus; for, as follows from the points of Section 4.2, it is this that is denied by non-sampling theorists (e.g., those accepting the likelihood principle).[19]

The (after-trial) uses of error probabilities are many, but they may all be traced to a single source: *the fact that error probabilities are properties of the procedure that generated the experimental result.*[20] This permits error probability information to be used in getting the result to answer questions about the process that produced it, and so to answer questions about causes. For example, error probability considerations are valuable Pearson explains

> because *it helps us to assess the extent of purely chance fluctuations that are possible* . . . . the result of applying the statistical test with its answer in terms of the chance of a mistaken conclusion if a certain rule of inference were followed, will help to determine the lines of further experimental work and the degree of confidence with which we proceed provisionally to adopt a new technique. (Pearson 1947, pp. 176–7; emphasis added)

Let us reconsider Pearson's naval shell example. The (after-trial) question being asked was '[H]ow much of the difference between (i) two failures out of twelve and (ii) five failures out of eight is likely to be due to this inevitable variability' (Ibid., p. 171)? It is asked by testing hypothesis H:

> H: The observed difference is due to inevitable or 'chance' variability.

(Alternative J would assert it is due to a systematic difference in the two shells, with respect to successfully piercing the plate (see Note 8).) The statistic D is the difference between the proportions of successful perforations of the plate from the two types of shell. The sampling distribution of D tells us that the observed difference D_{obs} is one improbably far from what would be expected were H correct. (The difference falls in the rejection region of a test of size approximately

0.05.) Even if no repetitions are planned, this informs us of the origin of this difference, and there are many ways to express this information in assessing H. One is that the observed difference (in piercing ability) is not easily accounted for by inevitable variability in the shells and measurement procedures, indicating that it is due to some systematic difference in the processes. A second, Pearson's, is that the result "would raise considerable doubts as to whether the performance of the first type of shell was as good as that of the second" (ibid., p. 192; see also Note 9). Further information may be obtained from the sampling distribution of D over alternatives to H, e.g., to find how large a systematic difference would be needed to generate differences as large as D_{obs} fairly frequently.

I will say more about these causal interpretations later. Here I want to draw a contrast with approaches that do not make use of sampling distributions. From the work of Birnbaum, Armitage, and others, it is known that a likelihood evaluation can result in rejecting (or in some way disfavoring) a hypothesis H with high probability (in extreme cases with probability 1) even though H is true! So, one clearly cannot say that it is very improbable for H to be erroneously rejected in favor of J on a likelihood evaluation: it may often infer J, e.g., the observed difference is a systematic effect, when it is really due to inevitable fluctuations as hypothesized in H. That is, one cannot guarantee the low Type I error afforded by NPT tests. One can intuitively see how this may result from certain rules for (a) data generation, e.g., sampling until H is specifiably less likely than some alternative J,[21] as well as (b) hypothesis selection, e.g., construct for the test hypothesis one that makes the observed result maximally likely.[22] Since such procedures do not affect the likelihoods, they do not alter the input from the data, according to the likelihood principle, a fact known as 'irrelevance of the sampling plan'. Having described an extreme example in which "misleading interpretations will be suggested by the likelihood principle with probability unity", Birnbaum[23] concludes:

> Thus it seems that the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations. (Birnbaum 1969, p. 128)

Such error control is important for correctly identifying experimental effects. The naval shell example imagined no actual repetitions, but we are often interested in learning about repeatable effects, and here error probabilities play a further role – that of helping to correct errors. For example, if H is rejected by a difference extremely improbable if H is true, then, if we are wrong, we would find we were rarely able to get the difference to recur, and in this way discover our original error. With baised procedures, in contrast, one can persist in rejecting H erroneously. A remark from the statistician Lucien LeCam capsulizes the contrast I am drawing in this section:

> One of the claims [of neo-Bayesians] is that the experiment matters little, what matters is the likelihood function after experimentation . . . . it tends to undo what classical statisticians have been preaching for many years: think about your experiment, design it as best you can to answer specific questions, take all sorts of precautions against selection bias and your subconscious prejudices. It is only at the design stage that the statistician can help you. (Lecam 1977, p. 158)

To be sure, Bayesians (and likelihoodists, if they consider prior likelihoods[24]) can respond that they may avoid such erroneous and ad hoc inferences through suitable assignments to the priors by which they will multiply the likelihoods; that is, although bias in the data generation and hypotheses selection may leave likelihoods unchanged (and therefore will not affect the contribution of the data), a Bayesian can compensate in the prior probabilities of hypotheses. To this I imagine Pearson's reply to be: "Yes, you *can* avoid such errors, but it will depend on your having appropriate priors; while our aim is to ensure the avoidance of such errors *regardless* of prior beliefs".[25] About (subjective) Bayesian priors, Pearson has this to say:

> It seems to me that . . . [even with no additional knowledge] . . . I might quote at intervals widely different Bayesian probabilities for the same set of states, simply because I should be attempting what would be for me impossible and resorting to guesswork. It is difficult to see how the matter could be put to experimental test.
>
> . . . can it really lead to my own clear thinking to put at the very foundation of the mathematical structure used in acquiring knowledge, functions about whose form I have often such imprecise ideas? (Pearson 1962, pp. 278–79)

To summarize this section, NPT methods do, while non-sampling methods (based on the likelihood principle) do not, control error probabilities of tests, and for a Pearsonian, an inability to control error probabilities matters (in a scientific context) not because of the desire to avoid too often making erroneous inferences in some long run, but because of the desire to distinguish genuine from spurious effects in a given experiment. It is this aim – to correctly identify causes – that accords the importance to preliminary planning of which Pearson spoke.

4.4. *Pearsonian Logic for Learning about Causes*

To spell out the full Pearsonian 'logic' by which tests may be used to learn about causes goes beyond the present topic,[26] but I think the points just made suggest the main outlines for a causal interpretation of the components of NPT tests. The hypotheses are assertions about the possible causal origins of experimental outcomes. The tests are used, not to reach decisions nor to assign hypotheses degrees of support or probability, but to learn how discrepant a hypothesized cause is (or is not) from the actual cause. The justification for using tests with specifiably low error probabilities is the corresponding intuition about what justifies inferences about causes: that an experimental result warrants inferring the cause hypothesized by H if it is practically impossible for it to have arisen from alternative sources. Conversely, such an inference is not warranted if it is easy (probable) for the result to have arisen from sources other than H. In this way tests with good error properties can be made to coincide with those appropriate for learning about causes. Of course a test's appropriateness will depend on what one wishes to learn – something no theory of statistics can decide. But, even if a test is inappropriate for a given inquiry, knowledge of the error probabilities (even if only approximate) will at least enable one to criticize the test and find out what has (not) been learned from the result. We have discussed the suggested causal interpretation of a rejection of H in the case of the naval shell example. Pearson also gives clues toward interpreting failures to reject, i.e., examples where the test 'accepts' hypothesis H. Here, too, error probabilities afford answers to causal questions in just the way we ordinarily make use of knowledge of the severity[27] of a test to interpret what that test indicates about whatever it is being probed. The main points can be brought out in a simple example where the 'test' is an ultrasound probe.

Imagine an ultrasound probe being used to learn the extent of disease or lack of disease in a patient's artery, as measured by some mean quantity imaged, μ. Say μ_N is the value of this mean in a normal artery, and the further from this value the more diseased. The hypotheses H and J corresponding to one type of causal question would be:

H: $\mu = \mu_N$ (or $\mu \leqslant \mu_N$): The result is due to a normal artery
J: $\mu > \mu_{N:}$: The result is due to a diseased artery

Let the observed result, x_{obs}, lead to 'accept H'. That is, suppose x_{obs} is not improbably far from what results from normal arteries (i.e., where $\mu = \mu_N$). The result, in other words, is a score that allows H to 'pass' the test, so the report comes out that the patient is normal. The patient, let us suppose, wishes to know whether the result is really due to being normal. It seems obvious that knowledge of how probable an outcome such as hers would be under various assumptions about μ is relevant. That is what error probabilities tell us. Were the patient to find out, for example, that such a passing score would often result even with abnormal values of the quantity, she has grounds to deny that her result is a good indication that her mean is normal. However, if it is practically impossible for such a passing result to occur if her artery were abnormal, then it is a good indication that she is in the normal range. Again, that is what error probabilities reveal. Passing a test T of hypothesis H with score x indicates H to the extent that such a passing score is improbable were H false and some alternative true. This calls for calculating, not just the usual power function but,

Pr(such a passing score |J)

for J ranging over the alternative parameter space (i.e., for abnormal values of μ). This interpretation (of a passing result in this type of test) may be summarized in the following rule:

RULE 1: A passing result x_{obs} with a test T indicates the actual mean μ is less than some μ' just to the extent that a more extreme result (on the test statistic X) would, with high probability have occurred in test T if in fact $\mu \geqslant \mu'$.

That is,

RULE 1: x_{obs} with a test T indicates $\mu < \mu'$ just to the extent that $\Pr(X \geqslant x_{obs} | \mu')$ is high.

The term "indicates" here means "indicates as the source or cause of the result". Rejections admit of analogous rules.

This interpretation involves two calculations that differ from those in the NPT decision model. First, it is sensitive to the particular observed difference x_{obs}. Second, it involves calculating the $\Pr(X \geqslant x_{obs} | J)$ where

J ranges over alternative values of μ. In contrast, the NPT power function is defined as $\Pr(X \geqslant x^* \mid J)$ where x^* is the critical boundary beyond which the result is taken to reject H.[28] Still, this seems to reflect Pearson's use of tests, as well as much day to day use of NPT methods.

In any substantive causal inquiry, NPT methods would need to be used for a series of tests, aimed at rejecting each type of alternative to H. Rejecting a 'chance' hypothesis H, with its indication that some systematic factor is operating, is likely to be just a first step. Ruling out other substantive factors must be accomplished with subsequent statistical tests. As Pearson stressed, there is no need to justify any single test as best; several tests may be used to learn the answers to different questions, as well as to check each other's assumptions. It is only by spelling out how NPT affords this type of piecemeal approach that one can capture the use of these methods that I believe Pearson had in mind. Philosophers in search of a single global procedure for using data to update inferences have missed the power of a system which breaks down complex inquiries into manageable parts, and in which one question may be asked at a time.

5. CONCLUSION

I have argued that E. S. Pearson rejected the basic tenets of the decision philosophy that has come to be associated with NPT methods. It is to be hoped that criticisms of NPT which are really directed at that philosophy reconsider Pearsonian ideas of the use and rationale of tests. Ideally, the NPT methods, so widely used, will come to incorporate explicitly the piecemeal uses and causal interpretations implicitly used by many practitioners. Alterations in the key concepts may be required. For example, I suggest that it is useful to calculate (after-the-trial) the probability of a difference as significant as the one observed given various hypotheses. But we still retain what is central to sampling theory: the focus on a procedure's error probabilities. Such developments are entirely in keeping with Pearson's philosophy:

> There is perhaps in current literature a tendency to speak of the Neyman–Pearson contributions as some static system, rather than as part of the historical process of development of thought on statistical theory which is and will always go on. (Pearson 1962, p. 276)

NOTES

* A portion of this research was carried out during tenure of a National Endowment for the Humanities Summer Stipend Fellowship; I gratefully acknowledge that support. A version of this paper was presented at the 1987 meeting of the Society for Exact Philosophy. This paper benefited from discussions and communications with George Barnard and Isaac Levi. I thank Harlan Miller for helpful comments on earlier drafts.

[1] Karl Pearson's subjectivist philosophy contrasts with that of his son Egon.

[2] For consistency with my notation, I substitute H and J for his H_1 and H_2, respectively.

[3] Birnbaum's system, incomplete at the time of his death, sought to make explicit the correspondence between an NPT result and a statement about strength of evidence (e.g., conclusive, very strong, weak or worthless). For example, he interprets reject H against J with error probabilities α, β equal to 0.01 and 0.2, respectively, as very strong statistical evidence for J as against H. The main shortcoming, as I see it, is that 'Pearsonian' reasoning seems to require a system in which tests with the same α, β may yield results with very different amounts of evidential import. Birnbaum's rules do not seem to reflect such differences. Further criticism along these lines occurs in Pratt (1977). Attempts at 'evidential' interpretations of NPT are discussed more generally in Mayo (1985).

[4] Neyman's sentiment has been differently understood. It might help to record his next sentence:

> In fact, no test can reveal any definite information about any statistical hypothesis if the values of the observable random variables which are possible under this hypothesis are also possible under some alternative one. (Neyman 1952, p. 66)

[5] The point here is that since the prespecified error probabilities are identical, they do not help in discriminating these two results, which is one of the sources of criticisms of NPT. Other uses of error probabilities *can* make this discrimination, in particular, those which I suggest are involved in a 'Pearsonian' model of tests. An analogous problem arises for NPT confidence intervals.

[6] George Barnard, in private communication, explains his part in Fisher's reception of NPT. While he brought to Fisher's attention how the testing framework favored by Neyman turned tests into pragmatic-decision tools, Barnard also distinguished this from Pearson's philosophy. Barnard (1985) provides an excellent discussion of historical developments in statistics, Barnard's past and most recent contributions, as well as comments from a number of statisticians.

[7] Pearson follows this naval shell example through a number of papers. Pearson was directly involved in the statistical assessment of army weapons in World War II, and after.

[8] The first treatment falls under what Pearson calls Problem I, (Barnard's "2 × 2 independence trial" the question being restricted to just the twenty shells observed, the total number of failures being fixed at the observed one, seven. The test asks whether the observed difference is due to a random partition of the twenty individual shells, of which seven would fail to perforate in whichever group they are randomly included. The second way of treating this case views samples from the two processes as random samples from two populations, so the failure rates can vary from 0–12 and 0–8, respectively. The test

asks whether the probability of failure is the same in both. This falls under what Pearson calls Problem II (Barnard's "2×2 comparative trial"). For the naval shell example, Pearson deems the latter treatment, preferred by Barnard, more artificial than the former: it regards the experiment as though "it were made on twenty shells, to twelve of which has been randomly assigned the label 'Made by firm X' and to the other eight, 'Made by firm Y'" (Pearson 1947, p. 192). The question of which of a number of ways to treat the 2×2 case had been much debated by Barnard and Fisher at that time. Pearson's answer is that the appropriate sample space is given "by the nature of the random process actually used in the collection of the data". But armed with an understanding of tests, he does not think one must rigidly choose from among several plausible tests.

[9] Pearson's conclusion inadvertently switches the observation to 2 of 12 and 5 of 8 successful perforations, where originally they had been failures. So, the conclusion, consistent with the original statement, should raise doubts as to whether the second type of shell is as good as the first, rather than conversely.

[10] Here Pearson calls it the 'experimental probability set'.

[11] Where this is not achievable (e.g., certain tests with discrete probability distributions) the test can associate with each contour an upper limit to this error probability.

[12] In a mixed test certain outcomes instruct one to apply a certain chance mechanism and accept or reject H according to the result. Because long-run error rates may be improved using some mixed tests, it is hard to see how a strict follower of NPT (where the lower the error probabilities the better the test) can inveigh against them. This is not the case for one who rejects the decision model of NPT as Pearson does. A Pearsonian could rule out the problematic mixed tests as being at odds with the aim of using the data to learn about the causal mechanism operating in a given experiment. Ronald Giere has presented this type of argument against mixed tests, appealing to propensity notions. See, for example, Giere (1976).

[13] A notable exception is the exposition of tests in Kempthorne and Folks (1971) in which test statistics are explicitly framed in terms of distance measures. Their interpretation of tests shares other key aspects with the approach I am recommending. See note 28.

[14] Interestingly, Hacking (1972) raises such doubts in reviewing Edwards (1972) on the basis of the type of problematic cases discussed in the rest of this section. Hacking (1980) questions other aspects of his 1965 likelihood approach, allowing that NPT does provide an account of inference.

[15] Fisherian methods would also fall under sampling theory, as would some eclectic approaches.

[16] The likelihood principle falls out directly from Bayes' Theorem. Birnbaum is responsible for showing, to the surprise of many, that it follows from two other principles, sufficiency and conditionality (together, or conditionality by itself). This result – while greeted with dismay by many non-Bayesians, who balked at the likelihood principle, but had thought sufficiency and conditionality intuitively plausible (including Birnbaum himself) – was welcomed with open arms by Bayesians, who saw in it a new corridor to a key Bayesian tenet. A third way would be to steer a path between the likelihood principle and advocating any and all principles that decrease error probabilities, thereby keeping certain aspects of sufficiency and conditionality *when they are warranted.* Birnbaum, I take it, sought to articulate some such third way. Elsewhere, I attempt to utilize the 'Pearsonian philosophy' discussed here to carry this broad strategy further.

[17] In practice, however, at least some Bayesians find error probability considerations useful, but, for consistency, they need to give Bayesian justifications for their use. I. J.

Good provides the most systematic framework for linking sampling theory measures to Bayesian ones in his 'Bayes/non-Bayes' compromise. In connection to the discussion of this section, see Good (1981).

[18] It is not that non-sampling methods lack theories of experimental design. The contrast is that in sampling theories, the plan of data and hypotheses generation is reflected directly in assessing the import of the data: different plans produce different error probabilities.

[19] It has often been suggested, e.g., Hacking (1965), that error probabilities, while acceptable for before-trial planning, should be replaced with other measures (e.g., likelihoods) after the trial. Pearson takes up and rejects this same proposal, raised by Barnard in 1950, reasoning that

> if the planning . . . is based on a study of the power function of a test and then, having obtained our results, we do not follow the first rule but another, based on likelihoods, what is the meaning of the planning? (Pearson 1963, p. 228).

[20] It may be objected that there are different ways to model the procedure. That is correct, and this enables different but interrelated questions to be asked to great advantage. This relates to Pearson's rejection of routine uses of tests in Section 3.3.

[21] Armitage's (1961) example is of this sequential sort: the rule instructs one to stop when the likelihood of a (simple) hypothesis H reaches an arbitrarily small value α for the first time. The result is that H is guaranteed (i.e., with probability 1) an assignment of so small a likelihood, even if H is true. This type of case and why it vitiates Hacking's (1965) likelihood testing approach is discussed in Mayo (1981).

[22] For example, in considering a series of hypotheses about various characteristics C_1, $C_2, \ldots, C_n$ in some population, e.g., about the proportion that share characteristic C_i, a rule might choose only to test a hypothesis for which the data accords maximal likelihood. In cases where the after-trial specification of hypotheses vitiates pre-specified error probabilities, the NPT test insists upon predesignation of hypotheses. A good discussion of how this type of after-trial construction leads to bias in Bayesian estimation occurs in Giere (1969).

[23] Birnbaum's example considers hypotheses about the mean μ and standard deviation σ of a normal distribution. An observation x would accord maximum likelihood to the hypothesis H that $\mu = x$ and $\sigma = 0$. Let the true hypothesis be J that $\mu = 0$ and $\sigma = 1$. Then, with probability 1, a likelihood appraisal would erroneously favor H over J. Birnbaum attributes an analogous example to Neyman (1952).

[24] A. N. F. Edwards (1972) is perhaps the most thorough going likelihood approach incorporating prior likelihoods.

[25] Of course where the hypotheses are assertions to which a frequentist prior can be assigned, it is open to a sampling theorist to apply Bayes' theorem. Suppose we were doing an experiment involving a randomly-selected naval shell, where p% of such trials would yield a shell with a rate of successful perforations equal to r. In that circumstance a frequentist could sensibly assign a probability of p to a hypothesis "the rate of successful perforations of a naval shell equals r". This is not the typical circumstance, however, leading Neyman and Pearson to seek methods that do not require such priors.

[26] See Mayo (1985, 1988, 1991) for more detailed, but still incomplete, attempts.

[27] This notion, as I use it here, accords with Popper's general idea of a 'severe test'.

See, for example, Popper (1972, pp. 14, 353–54). I develop and apply a formal notion of severity in Mayo (1988 and 1991).

[28] Kempthorne and Folks (1971) erect a system based on this type of calculation. Inverting significance tests, an observed result is used to give a family of confidence, or what they call 'consonance', intervals – one for each value of the confidence level (alternatively, the significance level). Features of their approach helped shape the one I recommend. The sets of parameter values 'indicated' according to Rule 1 are formally equivalent to their consonance intervals (for certain values of α). However, there are differences in interpretation, some of which are discussed in Mayo (1985) (e.g., Note 13), as well as in the experimental uses to which I intend these calculations to be put.

REFERENCES

Armitage, P.: 1961, 'Consistency in Statistical Inference and Decision', *Journal of the Royal Statistical Society* **B23**, 1–37.

Barnard, G. A.: 1985, *A Coherent View of Statistical Inference*, Technical Report Series, University of Waterloo, Waterloo Ont.

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

Birnbaum, A.: 1969, 'Concepts of Statistical Evidence', in S. Morgenbesser, P. Suppes, and M. White (eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, St. Martin's Press, New York, pp. 112–143.

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**, 19–49.

Edwards, A. W. F.: 1972, *Likelihood*, Cambridge University Press, Cambridge.

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

Fisher, R. A.: 1955, 'Statistical Methods and Scientific Induction', *Journal of the Royal Statistical Society* **B17**, 69–78.

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

Giere, R. N.: 1976, 'Empirical Probability, Objective Statistical Methods, and Scientific Inquiry', in W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol. II, D. Reidel, Dordrecht, pp. 63–101.

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

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

Hacking, I.: 1972, 'Likelihood', *British Journal for the Philosophy of Science* **23**, 132–37.

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

Howson, C. and Urbach, P.: 1989, *Scientific Reasoning: The Bayesian Approach*, Open Court, La Salle.

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

Kyburg, H. E. Jr.: 1971, 'Probability and Informative Inference', in V. P. Godambe and

D. A. Sprott (eds.), *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, pp. 82–103.

Kyburg, H. E. Jr.: 1974, *The Logical Foundations of Statistical Inference*, D. Reidel, Dordrecht.

LeCam, L.: 1977, 'A Note on Metastatistics or 'An Essay Toward Stating a Problem in the Doctrine of Chances'', *Synthese* **36**, 133–60.

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

Lindley, D. V.: 1971, 'The Estimation of Many Parameters', in V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, pp. 435–47.

Lindley, D. V.: 1976, 'Bayesian Statistics', in W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol. II, D. Reidel, Dordrecht, pp. 353–62.

Mayo, D.: 1981, 'Testing Statistical Testing', in J. C. Pitt (ed.), *Philosophy in Economics*, D. Reidel, Dordrecht, pp. 175–203.

Mayo, D.: 1982, 'On After-Trial Criticisms of Neyman-Pearson Theory of Statistics', in P. Asquith and T. Nickles (eds.), *PSA 1982*, Vol. 1, PSA, East Lansing, pp. 145–58.

Mayo, D.: 1983, 'An Objective Theory of Statistical Testing', *Synthese* **57**, 297–340.

Mayo, D.: 1985, 'Behavioristic, Evidentialist, and Learning Models of Statistical Testing', *Philosophy of Science* **52**, 493–516.

Mayo, D.: 1988, 'Toward a More Objective Understanding of the Evidence of Carcinogenic Risk', in A. Fine and J. Leplin (eds.), *PSA 1988*, Vol. 2, PSA, East Lansing, pp. 489–503.

Mayo, D.: 1991, 'Novel Evidence and Severe Tests', *Philosophy of Sciences* **58** (December).

Neyman, J.: 1950, *First Course in Probability and Statistics*, Henry Holt, New York.

Neyman, J.: 1952, *Lectures and Conferences on Mathematical Statistics and Probability*, 2d edn., Graduate School of U.S. Department of Agriculture, Washington DC.

Neyman, J.: 1971, 'Foundations of Behavioristic Statistics', in V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, pp. 1–13 (Comments and Reply, pp. 14–19).

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**, 175–240 (reprinted in *Joint Statistical Papers*, University of California Press, Berkeley, 1967, pp. 1–66).

Neyman, J. and Pearson, E. S.: 1933, 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', in *Philosophical Transactions of the Royal Society* **A231**, 289–337 (reprinted in *Joint Statistical Papers*, University of California Press, Berkeley, 1967, pp. 203–39).

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*, University of California Press, Berkeley, 1966, pp. 169–97).

Pearson, E. S.: 1950, 'On Questions Raised by the Combination of Tests Based on Discontinuous Distributions', *Biometrika* **37**, 383–98 (reprinted in *The Selected Papers of E. S. Pearson*, University of California Press, Berkeley, 1966, pp. 217–32).

Pearson, E. S.: 1955, 'Statistical Concepts in Their Relation to Reality', *Journal of the Royal Statistical Society* **B17**, 204–07.

Pearson, E. S.: 1962, 'Some Thoughts on Statistical Inference', *Annals of Mathematical*

Statistics **33**, 394–403 (reprinted in *The Selected Papers of E. S. Pearson*, University of California Press, Berkeley, 1966, pp. 276–83).

Pearson, E. S. and Wilks, S.: 1933, 'Methods of Statistical Analysis Appropriate for k Samples of Two Variables', *Biometrika* **25**, 353–78 (reprinted in *The Selected Papers of E. S. Pearson*, University of California Press, Berkeley, 1966, pp. 81–106).

Popper, L.: 1972, *Objective Knowledge*, Oxford University Press, Oxford.

Pratt, J. W.: 1977, "'Decisions' as Statistical Evidence and Birnbaum's 'Confidence Concept'', *Synthese* **36**, 59–69.

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

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

Spielman, S.: 1972, 'A Reflection on the Neyman–Pearson Theory of Testing', *British Journal for the Philosophy of Science* **24**, 201–22.

Department of Philosophy
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
U.S.A.