On After-Trial Criticisms of Neyman-Pearson Theory of Statistics[1]

Deborah G. Mayo

Virginia Polytechnic Institute
and State University

1. Introduction

Whether it is due to the incompleteness of information, inaccuracies of measurement, or stochastic nature of phenomena, a great deal of scientific inference requires probabilistic considerations. In carrying out such inferences, the statistical methods predominantly used are from the Neyman-Pearson Theory of statistics (NPT). Nevertheless, NPT has been the target of such severe criticisms that nearly all philosophers of induction and statistics have rejected it as inadequate for statistical inference in science. If these criticisms do in fact demonstrate the inadequacy of NPT, then a good portion of statistical inference in science will lack justification. Because of the seriousness of such a conclusion, it is important to carefully consider whether critics of NPT have succeeded in demonstrating its inadequacy.

In this paper I attempt to (1) clarify what I take to be the key issues around which the major criticisms against NPT revolve; and (2) argue that such criticisms fail to provide grounds for rejecting NPT as inadequate for science. I begin by drawing a fundamental distinction between the conception of the aims of statistical inference underlying the NPT, and the conceptions underlying rival views. Corresponding to the distinction between conceptions of statistical inference is a fundamental distinction between the criteria appropriate for judging the adequacy of statistical theories. Clearly, what is adequate for accomplishing the aims of NPT need not be adequate for accomplishing a fundamentally distinct set of aims. I maintain that criticisms of NPT involve judging NPT on the basis of criteria fundamentally distinct from those appropriate for NPT. By showing that NPT fails to satisfy these criteria, I claim, they succeed only in showing that accomplishing the aims underlying *these* criteria is incompatible with accomplishing the aims underlying NPT criteria. Unless there is reason to think

PSA 1982, Volume 1, pp. 145-158

that this poses a problem for NPT, it does not warrant inferring that NPT is inadequate--whether the inadequacy is merely being too narrow in its applicability or, in the extreme case, being totally refuted.

That is, what I call a weak claim (WC): *NPT fails to satisfy criterion C*, only warrants inferring a strong claim (SC): *NPT is inadequate*, if one accepts the additional premise (P): *NPT is adequate only if it satisfies criterion C*, for some specified criterion C. To justify accepting (P) it must be shown either that (I) NPT claims to satisfy criterion C, or (II) NPT should satisfy criterion C. I argue that criticisms against NPT can be taken as providing grounds for (I) only by misconstruing the aims of NPT, and as grounds for (II), only by presupposing the correctness of a conception of statistics fundamentally alien to NPT, which effectively begs the question against it. Failing to substantiate the needed premise (P), I conclude, prevents criticisms of NPT from going beyond a weak claim (WC): that NPT fails to satisfy a given criterion C. I am not thereby claiming that such criticisms serve no purpose; they serve an important function in delimiting the aims which one can rightly expect NPT to accomplish. Nor do I claim to have shown that NPT really is adequate for science; this requires a positive argument showing the appropriateness of NPT aims. What I do claim to have shown is that the widespread rejection of NPT as inadequate by philosophers is premature; the inadequacy of NPT has not been demonstrated.

2. Error Probabilities vs. E-R Measures

The view considered most plausible by most philosophers of induction and statistics, is that a theory of statistical inference, like a theory of deductive inference, should serve to assess the relationship between evidential claims and conclusions. Since, in the statistical case, the relationship between evidence (i.e., data) and a given conclusion may be weaker than deductive validity, a theory of statistics, on this view, should seek to provide a way of measuring *degrees* of the evidential relationship between them. To this end, measures of *evidential-relationship* (E-R measures) have been developed. Theories of statistics based on such E-R measures may be referred to as *E-R theories*. Carnap's confirmation measure, (subjective) Bayesian measures of degrees of belief, Fisher's fiducial probabilities, Hacking's measure of support, and Kyburg's epistemological probabilities, are only a few of the many E-R measures upon which E-R theories have been based. A theory of statistics will be adequate for the aim of E-R theories to the extent that it provides (absolute or relative) E-R measures that adequately express the degree of the evidential strength that *specific* data affords *specific* claims of interest.

In contrast, NPT views the aim of a theory of statistics to be that of providing *general* procedures or rules for making statistical inferences; where these general procedures are guaranteed to have sufficiently low probabilities for leading to various erroneous inferences.[2] NPT inferences are not assertions about E-R measures, but about certain properties of a population of interest, called *parameters*. Such inferences are based on the result of an *experimental trial* consisting of taking a sample of the population and observing some property of this sample, called a *statistic*. For example, one may observe the proportion of 'heads' (a statistic) in a sample of tosses of a coin to make inferences about the proportion of 'heads' in the (hypothetical) population of all tosses of the coin (a parameter). It will facilitate our discussion to outline two central types of NPT inferences about a parameter θ on the basis of observing statistic S; *hypotheses tests*, and *confidence interval* (CI) *estimates*.

General NPT Procedures: A hypotheses testing procedure consists of a rule which specifies which of the possible values of statistic S are to result in rejecting a specified hypothesis H (about θ) in favor of some alternative hypothesis $\bar{H}$. These values form the *rejection region*, RR. Hence, a testing rule takes the form: Reject H iff S is in RR. A general rule for confidence interval (CI) estimation, which I refer to as a *CI estimator*, indicates the specific CI estimate that should result from *each possible* value of S. A typical CI estimator has this form: Estimate that the true value of θ is in the interval $[S - c, S + d]$, i.e., estimate that $(S - c \leq \theta \leq S + d)$.

Specific Inferences: Once the trial is carried out, and S is observed to have specific value s, these general rules yield specific inferences. For example, suppose s is in RR. Then the test concludes: Reject H in favor of $\bar{H}$. The CI estimator concludes: $(s - c \leq \theta \leq s + d)$.

NPT Criteria: The specific inferences that result from a rule will vary with different values of S; some erroneous, others correct. A testing rule is adequate only if it is possible to guarantee, *before* the trial is made, that (regardless of the true value of θ) the probability it will lead to erroneously rejecting H is no more than some appropriately small number, called the *size* of the test. In addition, it should be able to guarantee that it will correctly accept $\bar{H}$ with suitably high probability, called the *power* of the test. Similarly, a CI estimator is adequate only if it has an appropriately large probability for leading to correct CI estimates, called the *confidence level* (CL) of the estimator. Size, power, and CL's are examples of *error probabilities*, and NPT inferences are judged adequate on the basis of the error probabilities of the rules from which they are generated.

3. Before-Trial Criteria vs. After-Trial Criteria

Unlike E-R measures, error probabilities hold only for general inference rules *before-the-trial* is made. Given the frequency view of probability underlying NPT, a rule's error probability is the relative frequency with which it will lead to correct (or incorrect) inferences in a sequence of (similar or very different) applications of the rule. For example, if the CI estimator considered above has a CL equal to .95, we can assert before-the-trial that, for any value of θ, correct to substitute S with the observed value s in this assertion. This is clearly seen in the case where the result of the substitution is $P[(1 \leq \theta \leq 3)/\theta] = .95$, and $\theta = 5$. For, the probability that the resulting CI (i.e., $(1 \leq \theta \leq 3)$) is correct is 0, not .95. NPT is intended for inferences where the parameters of interest are to be treated as constants. Hence, a *specific* NPT inference is either correct or not; i.e., it is true either 100% of the time or 0% of the time. So, the only probability that a specific inference may have is one of the trivial ones; 1 or 0.

It follows that error probabilities do not serve to express the degrees of probability, support, confirmation or any other E-R measure, that may be assigned to specific inferences. It is not surprising, then, that inferences that are adequate according to NPT criteria about error probabilities may not be adequate according to criteria about E-R measures. But critics of NPT maintain that criteria about E-R measures are what matter in analyzing specific inferences, after-the-trial, i.e., *after-trial analysis*. As such, they conclude that NPT is inadequate for after-trial analysis. Such criticisms may be referred to as *after-trial criticisms of NPT*. The most serious criticisms of NPT tend to be cases of after-trial criticisms of the following form: First, (1) an E-R measure is selected as appropriately assessing the extent of the evidential strength that specific data affords specific claims. Second, (2) a criterion is set out, based on the E-R measure selected, for judging the adequacy of inferences after-the-trial; call it criterion C. Third, (3) an example is constructed in which an inference that is more satisfactory, according to NPT (before-trial) criteria about error probabilities, is less satisfactory according to (after-trial) criterion C.

From this one may infer what I have termed a weak claim (WC):

(WC): NPT fails to satisfy (after-trial) criterion C.

If, in addition, one holds premise (P):

(P): Satisfying (after-trial) criterion C is necessary for NPT to be adequate (for after-trial analysis),

then one may infer the strong claim (SC):

(SC): NPT is inadequate (for after-trial analysis).

The problem with after-trial criticisms arises only if they are taken as grounds for inferring (SC). I argue that no such inference is warranted since they fail to provide adequate grounds for accepting (P). It may be suggested that (P) is intuitively obvious, and hence, requires no justification, as in the following remarks of Cederic Smith: "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 fails to provide this. One must conclude that it is not an appropriate theory of inference." (Smith 1977, p. 74). Here, (P) amounts to requiring that NPT provide an (after-trial) E-R measure of probability (i.e., a *posterior probability*). But NPT is based on the premise that an adequate theory of inference need not satisfy such a requirement. And if a criticism of NPT is based on assuming that a basic NPT premise is false, then it is clearly begging the question against it. In what follows, I argue that after-trial criticisms of NPT are based on just this sort of assumption.

4. After-Trial Criticisms of NPT Hypotheses Tests

I base my argument upon those types of after-trial criticisms that are most serious, as well as most influential. As representatives of these, I take the criticisms raised by three philosophers: Hacking, Spielman, and Seidenfeld. Hacking's criticism served as a model for the other two, both of whom attempted to improve upon it. Hacking tried to show that NPT tests are "suitable for before-trial betting, but not for after-trial evaluation." (Hacking 1965, p. 99). After-trial evaluation, on his view, involves measuring the extent to which specific data s *supports* hypotheses of interest. The E-R measure chosen for this purpose is the *likelihood function* (LF). The LF of hypothesis H given specific data s is the probability (or in the continuous case, the density) of s given that H is true, i.e., P(s/H). Tests are judged on the basis of the following (after-trial) *criterion of support*. [CS]:

[CS]: A test should reject hypothesis H on the basis of specific data s iff there is a rival hypothesis $\bar{H}$ much better supported by s, as measured by the LF, i.e., a test should reject H on the basis of s iff there is an $\bar{H}$ such that P(s/$\bar{H}$) is much greater than P(s/H).

Hacking provides an example where a test that is better according to NPT (before-trial) error-probabilities, is much worse according to his (after-trial) criterion [CS]. More specifically, while the test has a higher probability of correctly accepting $\bar{H}$ in a sequence of trials (i.e., has a higher power); a specific instance of this sequence is seen to result in accepting H although H is false, and hence has *no* support. That is, the specific instance gives rise to an s which leads to accepting H although P (S/H) = 0. Similarly, a test which fails miserably on NPT criteria is seen to satisfy [CS]. From this Hacking concludes the strong claim (SC): NPT tests are inadequate for after-trial analysis. But his argument only provides grounds for the weak claim (WC): NPT tests fail to satisfy (after-trial)criterion [CS]. In one sense even (WC) may be questioned; for[3] Hacking's examples involve tests which are not "best" on NPT criteria; and best NPT tests, in his examples, do satisfy [CS]. However, other examples can be constructed that show the incompatibility between NPT error criteria and Hacking's criterion of support [CS], thus establishing (WC). But unless there are grounds for supposing (P) (NPT tests must satisfy [CS]) there is no warrant for inferring (SC) (NPT tests are inadequate after-the-trial).

Upon examining the examples showing the incompatibility of NPT criteria and [CS], I think it is clear that, rather than justifying the needed premise (P), they provide positive grounds for *denying* (P), and rejecting [CS] as an inadequate after-trial criterion. For, [CS] permits one to reject a hypothesis in favor of the most _ad hoc_ hypothesis, formulated (after-the-trial) to perfectly fit specific experimental result, s. The reason is that, even if the _ad hoc_ hypothesis is clearly false, [CS] directs one to accept it simply because it has the maximum value of the LF;and hence, is best "supported". For example, observing a coin to land 'heads' gives maximum support to the hypothesis that both faces of the coin are heads-even when this is known to be false. In this way [CS] frequently leads to erroneous inferences. Indeed, it can be shown (see Mayo 1981b) to yielda sequence of inferences, 100% of which are wrong, even in the simplest one parameter case! Birnbaum (1969) demonstrates this for the two parameter case. In fact, the reason Neyman and Pearson (e.g., Neyman 1952) explicitly reject an after-trial analysis based on LF's alone, is that they saw it would prevent the (before-trial) guarantees of low probabilities of errors which they sought.

Clearly, then, Hacking's criterion [CS] conflicts radically with the aims of NPT. So, merely assuming (P) (NPT must satisfy [CS]) is tantamount to assuming that NPT should abandon its own aims. And just such an assumption is necessary if Hacking's argument is to be taken as grounds for inferring (SC) (NPT tests are inadequate for after-trial analysis). Although Hacking does provide positive arguments in favor of his own theory of statistics based on [CS], this does not prevent his after-trial criticism of NPT from this question-begging assumption. For his criticism is based on assuming the superiority of his own theory--a theory to which NPT is radically opposed. Moreover, while Hacking has since abandoned his likelihood theory, his after-trial criticism of NPT has still been seen by many as providing the groundwork for a non-question begging demonstration of NPT's after-trial inadequacy.

Spielman (1973) claims that by reconstructing Hacking's criticism, he can "show that NPT is inadequate on its own terms" (p. 202) and so provide a genuine "refutation" of NPT tests. However, his after-trial criticism, I argue, is flawed in much the same way as Hacking's--despite his attempt to avoid such flaws. Like Hacking, Spielman wants to show that the error probabilities of NPT tests (i.e., size and power) are irrelevant "once an experiment is performed, and a decision that really counts has to be made" (p. 211), (i.e., for after-trial analysis). According to Spielman, an after-trial analysis of a specific inference, based on specific data s, requires a measure of its _reliability_; and the E-R measure he selects for this purpose is the (posterior) probability that the specific inference is correct. The criterion used in judging tests is the following (after-trial) _criterion of reliability_ [CR]:

[CR]: A test should lead to a specific testing inference (i.e., accept or reject) on the basis of specific data s, only if the inference is sufficiently reliable, as measured by the probability that it is correct.

Spielman considers an example of a NPT test which has appropriately high probabilities of yielding correct inferences in a sequence of applications of the testing rule (and hence satisfies NPT error probability criteria); but which cannot guarantee that specific instances of this sequence will yield inferences with high probabilities of being correct (and hence fails to satisfy his (after-trial) criterion [CR]). In other words, Spielman shows that a testing rule,based on observing statistic S,may have a high probability (before-the-trial) of leading to correct inferences; while the probability (after-the-trial) of a specific observed value, s, yielding a correct inference may not be high. This entails the weak claim (WC): NPT fails to satisfy (after-trial) criterion [CR]. But when Spielman goes on to infer the strong claim (SC) (NPT tests are inadequate) it becomes clear that his criticism of NPT is not "on its own terms" - contrary to what he had intended. Again, the problem involves substantiating premise (P) (NPT must satisfy [CR] in order to be adequate). For, as we saw in Section 3, NPT never intended to provide an (after-trial) E-R measure of reliability; and it is explicitly denied that error probabilities apply once a specific s is substituted for S. Nevertheless, Spielman maintains (I) that NPT does intend to satisfy his (after-trial) criterion [CR]; or, (II) if not, it should.

Spielman offers the following argument in support of (I). He claims (i) that it is "implicit in the conceptual framework of NPT" (p. 207) that error probabilities of NPT general procedures are intended to justify their specific applications. And since according to Spielman (ii) for specific inferences to be justified they must be sufficiently reliable (in the sense of having sufficiently high probabilities of being correct), he concludes (iii) that NPT error probabilities are intended to guarantee that specific NPT inferences are sufficiently reliable (in his sense). From this it follows (I) that NPT intends to satisfy his (after-trial) criterion [CR] (and so, (P) is true). But it has been shown that (WC): NPT fails to satisfy [CR]. So, since it fails to satisfy the criterion which it intends to satisfy, NPT is refuted as promised. However, his proposed refutation does not succeed; for his argument for (iii) is flawed.

The flaw is in assuming premise (ii). For NPT is based on denying that (ii) is the case. Rather, it holds that specific inferences get their justification from having arisen from general procedures with appropriate error probabilities (in a given long run sequence). Hence, in assuming (ii) Spielman is assuming the correctness of a criterion that radically conflicts with the aim of NPT. And in calculating degrees of probability of specific inferences to be other than 0 or 1, he makes use of probabilities that are invalid on the frequency view underlying NPT. Hence,(iii), and so (I) is false; and there is no longer any basis for regarding error probabilities as "dangerously misleading."(p. 202). Only by misinterpreting them as (after-trial) E-R measures of reliability do they appear misleading.

Spielman admits that if NPT is only interested in guaranteeing low error probabilities in long run sequences, then (I) is false and he has not succeeded in refuting NPT. Nevertheless, he maintains that having shown that (WC): NPT tests fail to satisfy his (after-trial) criterion of reliability [CR], "I have shown that [NPT] is too narrow to bother refuting." (p.214). But, he has not told us why failing to satisfy [CR] is a deficiency for NPT. That is, he has not shown (II) that NPT *should* satisfy [CR]. Yet in claiming that unless it does, then NPT is "too narrow to bother refuting," Spielman is presupposing (II); and is thereby begging the question against NPT. Graves (1978) provides a good discussion of these and other points concerning Spielman's after-trial criticism.

5. After-Trial Criticisms of NPT of Confidence Intervals (CI's)

Seidenfeld (1979) and (1981), like Spielman, sets out an after-trial criticism of NPT by reconstructing the argument given by Hacking. Seidenfeld's criticism shares the same thrust as those of Hacking and Spielman; namely, that (before-trial) error probabilities of NPT may fail to indicate the degree of support (Hacking), reliability (Spielman) or confidence (Seidenfeld) that should be assigned to specific inferences--as measured by an appropriate (after-trial) E-R measure. But, rather than direct his criticism at error probabilities of NPT tests, Seidenfeld directs it at error probabilities of NPT of confidence intervals (CI's); in particular, confidence levels (CL's). He claims he will show that "it seems reasonable to say, *before knowing the data*," that a CI estimator with a CL equal to p will lead to a correct CI estimate with probability p; "however, *having seen the value* x, it may be unreasonable to maintain the probability statement or use it to express a degree of confidence in the interval generated by the [CI estimator]." (Seidenfeld 1979, pp. 56-57).

But, as was noted in Section 3, (and as Neyman and Pearson[4] have repeatedly warned) applying a CL to a specific interval estimate leads to absurdities. For, parameter θ is viewed, by NPT, as a constant, and probabilities are viewed as frequencies in some sequence. Clearly, it makes no sense to say that the frequency with which θ is contained in a specific interval [a,b] is, say, .95. A specific CI: $(a \leq \theta \leq b)$ is either always true or always false; thus, the only probabilities it may be assigned are 0 and 1. So, to show that CL's fail to provide (after-trial) measures of probability, is just to show that they fail to provide a type of (after-trial) evaluation that is illegitimate from the point of view of NPT. This is precisely what the after-trial criticisms of NPT tests were seen to amount to.

In order to avoid just this flaw, Seidenfeld develops a clever strategy: He will mount his after-trial criticism without using any E-R measure that is illegitimate from the point of view of NPT. The only probabilities, compatible with the frequency view of probability, that can be assigned to specific estimates are 1 and 0--according to whether it is known to be correct or not. Interval estimates that are known to be correct (i.e., known to contain the true value of θ)are referred to by Seidenfeld as *trivial intervals*. Seidenfeld notes that "even on Neyman's conception of probability there is an acceptable probability for the trivial intervals. They carry a known probability 1." (Seidenfeld 1981, p. 283). Hence, by basing his criticism on trivial intervals, he hopes to carry out his strategy.

Seidenfeld judges NPT of CI's on the basis of the following (after-trial) *criterion of triviality* [CT]:

[CT]: A CI estimator should not yield specific CI estimates with CL's that conflict with their known probabilities. In particular, a EI extimator with CL less than 1 should not yield trivial CI estimates (i.e., estimates known to have probability 1.)

Seidenfeld considers an example where the CI estimator that is "best"[5], according to NPT (before-trial) criteria, is inferior to one deemed less than best by NPT, according to his (after-trial) criterion [CT]. Its inferiority lies in the fact that it yields trivial intervals more often, while sharing the same CL of .95; that is, it yields estimates whose CL (i.e., .95) more often conflicts with the known probability that the estimate is correct (i.e., 1). From this Seidenfeld infers the weak claim (WC): NPT of CI's fails to satisfy (after-trial) criterion [CT].

Even this inference may be questioned, it seems; for it is arguable that, in the example Seidenfeld considers, NPT would actually recommend, not the estimator Seidenfeld criticizes, but an alternative estimator which does not yield any trivial intervals; and hence, satisfies [CT]. I argue this in detail in Mayo(1981a).[6] Admittedly, the basis for recommending the alternative interval involves *informal* criteria about informativeness. So, Seidenfeld's example can still be taken to show that satisfying the *formal* NPT criteria need not result in satisfying criterion [CT]; and I assume this is all the weak claim (WC) is meant to assert. The real problem arises when Seidenfeld's argument is taken as grounds for going beyond (WC), and inferring the strong claim (SC): NPT of CI's is inadequate for after-trial analysis.

At some points, Seidenfeld suggests that he only intends to show (WC); for his only concern is to show "that the N-P theory cannot serve as an adequate replacement for an inductive logic" (Seidenfeld 1979, p. 37), where an "inductive logic" is taken to require some (after-trial) E-R measure. And having shown (WC), that NPT of CI's fails to satisfy [CT], he has shown that CL's fail to provide valid E-R measures of probability or confidence. For, (WT) entails that an estimate in which one has 100% confidence may have a CL less than 100%. Nevertheless, Seidenfeld's concern with using only an E-R measure that is valid for NPT clearly suggests that he intends his argument to show some genuine flaw within NPT itself. He claims that "it is my goal in part, to strengthen Hacking's evaluation by showing that N-P best tests lead to clearly inferior confidence intervals, based on [the NPT criteria of "bestness"] alone." (p. 49). And this clearly implies that his goal is to show (SC): NPT of CI's is inadequate. However, these CI's are judged "clearly inferior" only in that they fail to satisfy Seidenfeld's criterion [CT]. To accomplish the goal of showing NPT of CI's leads to CI's that are "clearly inferior" on the basis of NPT criteria, he would have to justify premise (P): Failure to satisfy (after-trial) criterion [CT] renders NPT of CI's inadequate. I will argue that Seidenfeld's argument entails (P) only if NPT criteria are either misconstrued or rejected.

To justify (P), it must be shown either that (I): NPT claims to satisfy [CT]; or, (II) if not, it should. In support of (I), Seidenfeld appears to reason as follows: Since Neyman suggests that assigning an (after-trial) probability to a specific estimate is the "theoretically perfect solution" (Neyman 1937, p. 258), it would seem that when an (after-trial) probability is known--as in the case of trivial intervals, Neyman would want the NPT of CI's to provide it. Hence, Neyman would want to assign probability 1, and not a probability less than 1, to trivial intervals, i.e., he would want to satisfy criterion [CT]. It follows that (I).

Firstly, what Neyman personally would want is not the same as what NPT is logically capable of. NPT is intended to provide an adequate theory of statistics without (after-trial) E-R measures. While assigning (after-trial) probability 1 to trivial intervals is compatibl with the frequency theory of probability; such assignments are not strictly a part of NPT of CI's. Neyman specifically notes that within the theory of CI's, "we have decided not to consider [them]" (Neyman 1937, p. 263); for it is simpler to just assert the trivial interval itself. According to Seidenfeld, the problem with CI estimators failing to satisfy his criterion [CT], is "the tension between the confidence level (less than 100%) and a known probability (of exactly 100%)." (Seidenfeld 1981, p. 282). But, there is no such "tension" from the point of view of NPT. For, CL's always refer to the (before-trial) probabilities that CI estimators will lead to errors in a sequence of applications. And there is nothing contradictory, or even problematic, about having a specific estimate, which is known to be true (i.e., a trivial estimate), arise from a general estimating procedure which is known to lead to correct inferences less than 100% of the time (i.e., its CL is less than 1). It appears problematic only by misinterpreting CL's as providing (after-trial) E-R measures of confidence or probability. Seidenfeld's claim that failure to satisfy [CT] leads to trivial intervals being asserted "at strictly less than 100% confidence" (p. 281) involves such a misinterpretation. For the result of applying a CI estimator is just an assertion that θ is in the specific interval formed. Nothing is said about the degree of confidence which is to be attached to this assertion. Admittedly, the word 'confidence' encourages this sort of misinterpretation--but this is not a problem for NPT when it is correctly interpreted. Since CL's are not intended to provide (after-trial) measures of probability, even if such a probability is known; it follows that (I) is false.

Nor can Seidenfeld's argument be taken as grounds for inferring that (II): NPT should satisfy criterion [CT]. For, his own example shows that if the aim is to satisfy his criterion [CT], the result will be to recommend CI's that fail miserably on NPT criteria. And a CI will fail to be adequate, for the type of after-trial analysis that NPT is interested in, unless it satisfies these criteria. Hence, although the E-R measure Seidenfeld uses is legitimate on the frequency view, the after-trial criterion [CT] upon which he judges NPT of CI's involves after-trial considerations that are incompatible with the view of after-trial analysis underlying NPT. Failing to provide independent grounds for the correctness of his criterion [CT] prevents Seidenfeld's argument from legitimately showing (II). To merely assume (II) is tantamount to assuming that the aims of NPT should be rejected in favor of the aim underlying E-R theories; namely, to provide an after-trial measure of the evidential strength that specific data affords specific conclusions. Without grounds for either (I) or (II), premise (P) is unsubstantiated. Hence, Seidenfeld's argument can not be taken as grounds for inferring (SC): NPT of CI's is inadequate (for after-trial analysis). Moreover, Seidenfeld's criticism may be seen to follow the pattern of argument found in criticisms raised by Fisher (1956), Jaynes (1968), and Lindley (1971); and each, I claim, involves the same kind of problem.[7]

6. Conclusion

Each of the after-trial criticisms of NPT has been seen to involve judging statistical inferences on the basis of an (after-trial) criterion which reflects a very different view of the aim of statistical inference than the one embodied in NPT. For, underlying each criterion is the view that a theory of statistics should provide an expression of the extent of the evidential strength that specific data affords specific conclusions, after the trial is made. While such criteria are appropriate for judging E-R theories, they are inappropriate for judging NPT. Moreover, it has been shown that, unless NPT criteria are either misunderstood or rejected, the criticisms cannot be taken as grounds for thinking that it is necessary for NPT to satisfy these (after-trial) E-R criteria. In effect, the criticisms merely show the difference between NPT criteria and the (after-trial) criteria based on E-R measures. Therefore, it can be concluded that the after-trial criticisms of NPT fail to demonstrate the inadequacy of NPT.

Notes

[1] I would like to thank Ronald Giere and Isaac Levi for very helpful comments on an earlier draft of this paper. I am also grateful to Henry Kyburg and Teddy Seidenfeld for sharing their responses to a former paper (Mayo 1981a) with me; they were extremely useful in clarifying the key points of contention underlying the present paper.

[2] The distinction I draw between the NPT conception of statistics and the conception underlying E-R theories is parallel to the distinction drawn by Giere (1977) between testing and information models.

[3] A "best" NPT test of a given size is one which also has the maximum power of any other test with the same size (for the given hypotheses under test).

[4] See especially Neyman (1937, pp. 261-273; 1952, pp. 210-214; and 1977, pp. 118-119).

[5] A CI estimator, CI*, with a CL of p, is "best" according to the NPT formal criteria if there is no other estimator (for the given estimation problem) that also has a CL of p, but which has a smaller probability of yielding estimates containing incorrect parameter values than CI*.

[6] My argument, briefly, is this: Seidenfeld is able to present a NPT "best" CI estimator, with a CL of .95, that generates trivial estimates only by considering a case where parameter θ is known to have a specific upper bound. This is called the _truncated case_. Since the formal NPT criteria are not affected by this truncation, Seidenfeld concludes that NPT still recommends the same interval in the truncated case. I argue that by making use of the additional available information in this case, an alternative interval recommends itself; and this alternative CI does not give rise to trivial intervals.

My basis for claiming that, in the truncated case, this alternative CI is superior on the basis of NPT principles is _not_ that it satisfies criterion [CT]--for, NPT does not seek to do so. Rather, I argue, that NPT recommends the best CI estimator, _relative to the type of information_ in which one is interested; and the alternative CI yields more appropriate information in the truncated case.

[7] Their arguments are, roughly, the following. Before the trial, a CI estimator with CL equal to p has a probability of yielding correct CI estimates equal to p, in a sequence of applications of the estimator. However, after the trial yields specific data s, s may be seen as a member of a subset, T, of the original sequence of applications; and the probability that the CI estimator will yield a correct estimate _in this subset_ of applications may be known to be q, where $p < q$. It is argued that, in such cases, the correct CL to assign a specific estimate based on s is not p, but q. (In Seidenfeld's example q equals 1). It is concluded that NPT of CI's are inadequate.

But, unless it is assumed that CL's must provide (after-trial) E-R measures, there are no grounds for this conclusion. For, it is perfectly valid to assign a CL of p to estimates based on s. It would only be appropriate to assign it a CL of q _if_ it had been decided before the trial to limit the sequence of applications to those in T. It is up to the experimenter to decide, before the trial, which sequence of applications is appropriate for evaluating a given inference; it need not include all possible applications. However, once it is specified, the CL is fixed; it cannot be altered by the specific experimental result.

References

Barnett, V. (1973). Comparative Statistical Inference. New York: Wiley.

Birnbaum, A. (1969). "Concepts of Statistical Evidence." In Philosophy, Science and Method. Edited by S. Morgenbesser, et al. New York: St. Martins. Pages 112-143.

----------. (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.

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

Fisher, R.A. (1956). Statistical Methods and Scientific Inference. New York: Hafner.

Giere, R.N. (1976). "Empirical Probability, Objective Statistical Methods and Scientific Inquiry." In Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science, Volume II. (University of Western Ontario Series in Philosophy of Science, Volume 6.) Edited by W.L. Harper and C.A. Hooker. Dordrecht: Reidel. Pages 63-101.

----------. (1977). "Testing vs. Information Models of Statistical Inference." In Logic, Laws and Life. (University of Pittsburgh Series in the Philosophy of Science, Volume 6.) Edited by R.G. Colodny. Pages 19-70.

Godambe, V.P. and Sprott, D.A. (eds.). (1971). Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston of Canada.

Graves, S. (1978). "On the Neyman-Pearson Theory of Testing." British Journal for the Philosophy of Science 29: 1-23.

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

Jaynes, E.T. (1968). "Confidence Intervals vs. Bayesian Intervals." In Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Volume II. (University of Western Ontario Series in Philosophy of Science, Volume 6.) Edited by W.L. Harper and C.A. Hooker. Dordrecht: Reidel. Pages 175-213.

Lehmann, E.L. (1959). Testing Statistical Hypotheses. New York: Wiley.

Lindley, D.V. (1971). 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: 269-280.

-------. (1981b). "Testing Statistical Testing." In Philosophy in Economics. Edited by J.C. Pitt. Dordrecht: Reidel. Pages 175-203.

Neyman, J. (1935). "On the Problem of Confidence Intervals." Annals of Mathematical Statistics 6: 111-116.

---------. (1937). "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." Philosophical Transactions of the Royal Society of London Ser. A, 236: 333-380. (As reprinted in A Selection of Early Statistical Papers of J. Neyman. Berkeley and Los Angeles: University of California Press, 1967. Pages 250-290.)

---------. (1941). "Fiducial Argument and the Theory of Confidence Intervals." Biometrika 32: 128-150.

---------. (1952). Lectures and Conferences on Mathematical Statistics and Probability. 2nd ed. Washington: U.S. Department of Agriculture.

---------. (1977). "Frequentist Probability and Frequentist Statistics." Synthese 36: 97-131.

Pearson, E.S. (1962). "Some Thoughts on Statistical Inference." Annals of Mathematical Statistics 33: 394-403. (As reprinted in The Selected Papers of E.S. Pearson. Berkeley and Los Angeles: University of California Press, 1966. Pages 276-283.)

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

-------------. (1981). "On After Trial Properties of Best Neyman-Pearson Confidence Intervals." Philosophy of Science 48: 281-291.

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: 201-222.