11

The Test of Experiment: C. S. Peirce and E. S. Pearson

Deborah G. Mayo, Virginia Polytechnic Institute and State University

Peirce's philosophy of experimental testing shares a number of striking features with the Neyman and Pearson theory of statistics developed in the 1930s. For both, statistical methods provide, not means for assigning degrees of evidential support, confirmation, or belief to hypotheses on the basis of data, but procedures for testing (and estimating), whose rationale is their predesignated high frequencies of leading to correct results in some hypothetical long run. Isaac Levi is perhaps the first philosopher to have drawn out these key parallels between Peirce and Neyman-Pearson theory: "Peirce's inductions are inferences according to rules specified in advance of drawing the inferences where the properties of the rules which make the inferences good ones concern the probability of success in using the rules. These are features of the rules which followers of the Neyman-Pearson approach to confidence interval estimation would insist upon" (1980:138). Ian Hacking (1980) also discusses philosophical as well as historical connections between Peirce and Neyman and Pearson. Nor is it just the general view of statistical method that Peirce shares with Neyman and Pearson. One finds specific examples in Peirce that anticipate the mathematics of Neyman-Pearson hypothesis tests and confidence-interval methods. While tracing these mathematical ideas in Peirce is of interest in its own right, my focus in this chapter is to suggest what Peirce's experimental philosophy has to offer to contemporary debates between Neyman-Pearson and its rival statistical philosophies.

To get at the main issues, consider a Neyman-Pearson test of hypothesis *h*. The test is defined by a rule that indicates which of the possible experimental outcomes will be taken to reject *h* in favor of some alternative hypothesis *j*, and which will be taken to accept *h*. The cornerstone of Neyman-Pearson tests is their ability to ensure, *before the trial is made*, that (regardless of the true hypothesis) the rule will erroneously reject *h* and erroneously accept *h* no more than with some specifiably small probabilities. The probability of an erroneous rejection of *h*—the Type I error—and the probability of an erroneous acceptance of *h*—the Type II error—are the test's *error probabilities*. Error probabilities are not probabilities of hypotheses but the probabilities with which certain results would occur in some sequence of applications of such tests.

Despite the widespread use of Neyman-Pearson methods, most philosophers of statistics have questioned their relevance for statistical inference as opposed to the process of making routine decisions. A theory of statistical inference, it is typically thought, should provide measures of the *evidential strength* that data afford hypotheses. Examples of such evidential-strength measures are posterior probabilities and likelihood ratios. Let us call this view the evidential-strength view of statistical inference. Since Neyman-Pearson tests fail to provide measures of evidential strength, but only give error probabilities of procedures, they are inadequate for the evidential-strength view of statistical inference.

Neyman himself held a statistical philosophy that encouraged such criticisms. He continually denied that the tests provided methods for inductive inference, but rather construed tests as rules for *inductive behavior*. For Neyman (1950:258), "The 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*." These are not decisions to accept or believe that what is hypothesized is approximately true or not, Neyman stresses. Rather, "to accept a hypothesis means only to decide to take action *A* rather than action *B*" (ibid.:259). The criticisms of Neyman-Pearson theory are largely directed at Neyman's *inductive-behavior* model of the tests. The critics admit that the Neyman-Pearson test may be sensible in certain decision-theoretic contexts, the paradigm example being 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. 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, Neyman-Pearson tests can provide the desired guarantees. But the process of testing claims in a scientific context does not seem to have these features.

Although Neyman-Pearson tests are typically associated with this behavioral-decision interpretation, E. S. Pearson (not to be confused with his father, Karl), cofounder of the Neyman-Pearson methods, himself rejected it. In responding to Fisher's (1955) criticism of the term "inductive behavior," Pearson (1955:207) declared, "This is Professor Neyman's field rather than mine." Pearson stressed that, from the start, the Neyman-Pearson theory—with its introduction of the two types of error—reflected his view of statistical *inference*. According to him: "It was not until after the main lines of this theory had taken shape . . . 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" (Pearson 1955:205).

While Pearson continued to speak in his papers of using the tests inferentially, he admitted to never having described a systematic inferential model of Neyman-Pearson tests. The debate as to whether, and if so how, Neyman-Pearson theory is relevant for statistical inference continues. A key question is why information about a test's error probabilities is relevant to particular inferences. I suggest that a plausible answer is provided by the view of experimental tests found in Peirce and that Peirce's answer has much in common with, and even extends, the one Pearson suggests.

What Peirce provides is a philosophy of inductive inference that calls for precisely the sort of tools provided by the Neyman and Pearson theory. In describing his theory of inference, Peirce could be describing that of Neyman-Pearson.

> The theory here proposed does not assign any probability to the inductive or hypothetic conclusion, in the sense of undertaking to say how frequently that conclusion would be found true. It does not propose to look through all the possible universes, and say in what proportion of them a certain uniformity occurs; such a proceeding, were it possible, would be quite idle. The theory here presented only says how frequently, in this universe, the special form of induction or hypothesis would lead us right. The probability given by this theory is in every way different—in meaning, numerical value, and form—from that of those who would apply to ampliative inference the doctrine of inverse chances. (*CP* 2.748)

The supposition that the probability of the conclusion is needed, Peirce claims, stems from a faulty analogy with deductive inference. Since deductive inference tells us that, if such and such premises are true, then a given conclusion is true, it might be thought that inductive inference should tell us that, if such and such premises are true, then a given conclusion is probable. Peirce denies this, claiming, "In the case of analytic inference we know the probability of our conclusion (if the premisses are true), but in the case of synthetic inferences we only know the degree of trustworthiness of our proceeding" (*CP* 2.693).

Peirce suggests that if probability had always been understood as relative frequency, it would not have been thought that a final probability was wanted. Since the conclusion concerns a hypothesis that either is or is not true about this world, the only probabilities that a frequentist could assign to it are the trivial ones—1 or 0. What we really want to know, according to Peirce, is this:

> Given a certain state of things, required to know what proportion of all synthetic inferences relating to it will be true within a given degree of approximation. Now, there is no difficulty about this problem (except for its mathematical complication); it has been much studied, and the answer is perfectly well known. And is not this, after all, what we want to know much rather than the other? (*CP* 2.686)

The current debate in philosophy of statistics, as in Peirce's day, reflects opposing answers to Peirce's question—Neyman-Pearson theorists agreeing, Bayesians and others disagreeing. This opposition is revealed in what each considers to be relevant information for inference.

A theory based on error probabilities—in Peirce's words, on "how frequently, in this universe, the special form of induction of hypothesis would lead us right" (*CP* 2.748)—is interested not in just the observed result but in other possible outcomes that could result from using the inference rule. That is, it is interested in the entire sample space, which is why the Neyman-Pearson theory is called a sampling theory. For example, the probability that a test commits a Type I error is the probability that the test would reject hypothesis *h* erroneously. It is a function of all the outcomes that would have led to rejecting *h*.

This points up the crucial difference between Neyman-Pearson tests and rival approaches which are nonsampling approaches, such as Bayesian or likelihood accounts: The Neyman-Pearson approach views the probability of outcomes that would have happened as crucially important in interpreting what actually did happen. For the nonsampling approaches, the relevant evidence contributed by the data is fully contained in the likelihood ratio actually obtained—a point formally expressed in the *Likelihood Principle*, which follows from Bayes's theorem. According to the Likelihood Principle, to quote from Dennis Lindley (1976:361):

> If we have two pieces of data x_1 and x_2 with the same likelihood function, [i.e., the probability of x_1 given h equals the probability of x_2 given h], the inferences about [h] from the two data sets should be the same. This is not usually true in the orthodox [Neyman-Pearson] theory, and its falsity in that theory is an example of its incoherence. . . . As [Harold] Jeffreys has said, *What has what might have happened, but did not, got to do with inferences from the experiment?* (Emphasis added)

From the Bayesian point of view, the interest in what might have occurred renders Neyman-Pearson tests "incoherent"; but from the Neyman-Pearson point of view, Bayesian (and other nonsampling) methods are unpalatable just because they ignore what the data generation procedure might have produced, once the data are in hand. This reflects the incompatibility between the former's interest in measures of evidential strength and the latter's interest in controlling error probabilities. To defend the inferential use of Neyman-Pearson tests requires a rationale for this interest in "what would have but did not happen"—correspondingly, in error probabilities. Peirce anticipates and even strengthens the rationale to be suggested by Pearson.

The concern with error probabilities for Pearson, like the concern with "the trustworthiness of the proceeding" for Peirce, is directly tied to their view that statistical method should closely link experimental design and data collection with subsequent inferences. Pearson claims: "But looking back I think it is clear why we regarded [the test's error probabilities] as more meaningful than the likelihood ratio [of the given result]. . . . 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" (Pearson 1966d:277–78).

Pearson railed against the tendency to see statistical inference as beginning only after "data are thrown at the statistician and he is asked to draw a conclusion" (ibid.:278). Analogously, Peirce criticized opposing inference philosophies, (the so-called conceptualists) for ignoring the manner of collecting data. In criticizing John Stuart Mill, for example, Peirce insists that "an induction, unlike a demonstration, does not rest solely upon the facts observed, but upon the manner in which those facts have been collected" (*CP* 2.766). Peirce introduces the term "quasi-experimentation" to include the whole process of generating and analyzing the data as well as using the data to test a hypothesis. And "this whole proceeding," Peirce declared, "I term Induction" (*CP* 7.115; see editor's note). For Peirce, the "true and worthy" task of logic is to "tell you how to proceed to form a plan of experimentation" (*CP* 7.59).

The reason error probabilities become relevant when one considers the "preliminary experimental planning" is this: by considering ahead of time the probability an experiment has of

detecting errors in hypotheses of interest, the statistical theory can then specify what type of and how large a sample would be needed to put hypotheses to an efficient test of experiment. For Peirce "induction begins with a question or theory" (*CP* 2.775).

> The next business in order is to commence deducing from it whatever experiential predictions are extremest and most unlikely, . . . in order to subject them to the test of experiment, and thus either quite refute the hypothesis or make such corrections of it as may be called for by the experiments; and the hypothesis must ultimately stand or fall by the result of such experiments. (*CP* 7.182, emphasis added)

Introducing the Type II error—an innovation which Pearson declares to be his own improvement on Fisherian tests[1]—gives precision to Peirce's testing conception. Requiring a small Type II error ensures that a result is taken to accept hypothesis *h* only if such a result is unlikely, if in fact *h* was (specifiably) false and a rival hypothesis was true.[2] Analogously, for Peirce:

> When we adopt a certain hypothesis, it is not alone because it will explain the observed facts, but also because the contrary hypothesis would probably lead to results contrary to those observed. So, when we make an induction, it is drawn not only because it explains the distribution of characters in the sample, but also because a different rule *would probably have led to the sample being other than it is*.[3] (*CP* 2.628, emphasis added)

This concern with the probability that the sample *would have been other* than it is in reasoning from the actual sample obtained, then, reflects the same inferential concerns in Peirce and Pearson, namely, the concern that an agreement between data and a hypothesis *h* be taken to accept *h* as approximately true only by being the result of a severe test of *h*—one with a high probability of having rejected the hypothesis, if false. And this, in turn, links to both their concerns with preliminary planning—it is only by appropriate procedures for generating the data and hypothesis that this severity requirement is met. (A fuller discussion of the account of severity I have in mind occurs in Mayo 1991.)

The aim of securing severe or powerful tests underlies the importance that Peirce and Neyman and Pearson place on predesignation of hypotheses. Predesignation requires that the hypothesis be specified before obtaining the data to be used in its test. Again this touches the heart of the contemporary controversy between Neyman-Pearson sampling and nonsampling philosophies, that is, between error-probability principles and the Likelihood Principle. For those who accept the Likelihood Principle, once the data are obtained, it is irrelevant for assessing their evidential import how they were selected, or whether the hypothesis was predesignated. (This is often called "irrelevance of the sampling rule.") For such matters do not alter likelihoods. But they do alter error probabilities. Just as Neyman-Pearson theorists insist, as against their rivals, on the relevance of predesignation, Peirce is highly critical of predesignation being "singularly overlooked by those who have treated of the logic of [induction]" (*CP* 2.738).

In addressing this issue, Peirce and Pearson consider a basic test where they want to argue from a property *P* shared by a proportion of some sample to a hypothesis that the observed regularity is genuine—that is, it may be expected to hold about as well in some population or class *M*. In other words, they want to deny that the observed regularity is merely accidental, or due to chance. As against violating predesignation, Pearson (1966b:127) cautions: "To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the [chance] hypothesis is true." In so doing, Pearson stresses, the question is changed, and therefore also the corresponding error probabilities of the test.

Peirce's points mirror Pearson's. Peirce insists that "the conclusion be confined strictly to the question. If the instances examined are found to be remarkable in any other respect than that for which they were selected, we can draw no inference of the present kind from that" (*CP* 2.784). The reason Peirce gives is the same as Pearson's.

> But suppose we were to draw our inferences without the predesignation of the character *P*; then we might in

> every case find some recondite character in which those instances would all agree. That, by the exercise of sufficient ingenuity, we should be sure to be able to do this, even if not a single other object of the class *M* possessed that character, is a matter of demonstration. For in geometry a curve may be drawn through any given series of points, without passing through any one of another given series of points. (*CP* 2.737)

That is, we can always find *some property or other* that all members of a sample share—even if no other members of the population do. Hence, if predesignation is violated, one cannot say that the hypothesis that the regularity is genuine has been subjected to a severe test. One cannot say that, were the regularity merely due to chance, the test would probably not have led to the observed regularity. On the contrary, it is highly probable that it would have led to observing some such regularity, even if it were only due to chance. In thus arguing against violating predesignation, Peirce also anticipates proofs by Neyman, and more recently by statisticians Birnbaum and Armitage, that violating predesignation permits tests which are wrong with high probability—in extreme cases, with probability 1! Thus, as Birnbaum (1969:128) concludes, methods based on the Likelihood Principle "cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations."

Neyman-Pearson tests are often wrongly taken to always require predesignation, when in fact it is called for only in certain types of cases—those in which violating predesignation conflicts with the goal of controlling the error probabilities. I find Peirce to be clearer than Neyman and Pearson as regards the necessity of qualifying the predesignation rule (e.g., *CP* 2.730). As I see it, Peirce's qualification amounts to describing circumstances where violating predesignation "doesn't matter much" to the reliability of the proceeding and thereby does not much hamper the ultimate goal of statistical inference.

This links with the centerpiece of Peirce's philosophy of experimental inference—his argument for the self-correcting rationale of induction. Peirce claims that "in the individual case in hand, if there is any error in the conclusion, that error will get corrected by simply persisting in the employment of the same method" (*CP* 2.781). As Levi (1980:138) stresses, however, "Peirce is not claiming that induction is self correcting in the sense that following an inductive rule will, in the messianic long run, reveal the true value [of the proportion with a given property]."

In the case of hypothesis testing, I suggest that Peirce's idea is that if a particular conclusion is wrong, subsequent severe (or highly powerful) tests will, with high probability, detect it. For example, in a good test, hypothesis *h* is rejected by outcomes that are extremely improbable if *h* is true. Then, if we are wrong in rejecting *h* (*h* is actually true), we would find we were rarely able to generate such outcomes; in this way, we would discover our original error.

But I want to suggest an even more localized conception of error correction in Peirce—one that is relevant even where no repetitions of an experiment are planned or are even possible. I suggest that Peirce valued considerations about the trustworthiness of a proceeding—or in the more modern terminology we have been using, considerations about a test's error probabilities—because they serve as tools for rooting out specific errors in a case of experimental reasoning. Examples of such errors are mistaking chance correlations for systematic ones, mistakes about the value of a parameter, and mistakes about a causal factor. I limit myself to just one of the many examples I find in Peirce: a test carried out by Dr. Lyon Playfair. It illustrates both a mistake resulting from violating predesignation as well as how, arguing from cases where error probabilities are sustained, the original mistake is corrected.

While considering Dr. Playfair "an accomplished reasoner," Peirce describes how he violates predesignation in testing a hypothesis about a regularity between the specific gravity of a metal and its atomic weight. Looking at the specific gravities of three forms of carbon, Peirce tells us, Playfair seeks out and discovers a formula connecting them: each is a root of the atomic weight of carbon, which is 12. Peirce describes the test Playfair carries out to judge if this regularity can be expected to hold generally for metals, showing that several alleged instances of the formula really involve modifications not specified in advance. If one limits the instances to those for which the formula is predesignate, only half of them satisfy Playfair's formula. Peirce reasons: "Having thus determined the ratio, we proceed to inquire whether an agreement half the time with the formula constitutes any special connection between the specific gravity and the atomic weight of a metalloid" (*CP* 2.738).

Peirce then subjects this to a test of experiment. The falsity of the hypothesis that there is a special connection is the claim that the observed agreement is "due to chance." Peirce asks: How often would such an agreement be found, even if it were due to chance? To answer this, Peirce introduces a hypothetical chance distribution by matching the specific gravity of a set of elements, not with its own atomic weight, but with the atomic weight of some other element, with which it is arbitrarily paired. For example, the specific gravity of carbon is compared to the atomic weight of iodine. Note that Peirce is not running more trials of Playfair's experiment but considering how many agreements with Playfair's formula would occur in a case designed so that such agreements could only be due to chance and, using this information about *what would be*, is arguing about the cause of the agreements actually found. Peirce finds about the same number of cases satisfying Playfair's formula in this chance pairing of elements as Playfair found in comparing specific gravities and atomic weights of a given element. Peirce concludes: "It thus appears that there is no more frequent agreement with Playfair's proposed law than what is due to chance" (*CP* 2.738). So Playfair was mistaken in thinking he had found a special or systematic connection. With procedures that violate predesignation of the formula, in contrast, one can persist in this error.

Pearson, like Peirce, denies that the importance of low error probabilities is a concern with ensuring a low frequency of erroneous judgments in some long run—at least, in typical scientific uses of tests. Rather, Pearson (1966a:172) suggests that "the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgment" in understanding the process in a particular case.[4] For example, a consideration of error probabilities is valuable, Pearson explains, because "it helps us to assess the extent of purely chance fluctuations that are possible" (ibid.:176). If purely chance fluctuations can easily account for an observed regularity, as Peirce shows in the case of Dr. Playfair, there is no reason for taking it as a sign of a genuine regularity.

It is important to consider one further way in which the manner Pearson advocates for using error probabilities sheds light on the manner in which induction may be seen as self-correcting according to Peirce. The behavioral-decision model, typically associated with Neyman-Pearson tests, encourages tests to be viewed as automatic methods, or "recipes," for testing claims, often leading to their misuse. However, Pearson specifically denied that the tests are to be used as automatic routines for testing claims, insisting that "no responsible statistician . . . would follow an automatic probability rule" (ibid.:192) in scientific contexts.[5] Likewise, Peirce's argument that induction is self-correcting should not be thought to be claiming that some automatic routine for detecting errors exists.[6] As Nicholas Rescher (1978:14) argues: "It cannot be said too emphatically that Peirce does not hold the ill-advised view that the method of science is 'self-corrective' in providing some sort of automatic, cookbook procedure for devising good theories to put in place of bad ones."

The view Peirce holds, I suggest, is that inductive methods—viewed as methods for severe testing—provide systematic (though not automatic) procedures whose aim is detecting and correcting key errors in experimental reasoning. The modern theory of Neyman-Pearson tests may be seen to provide Peirce's experimental philosophy with mathematically rigorous methods for carrying out this aim. It does so in two ways. First, Neyman-Pearson theory provides standard statistical models in which to couch hypotheses about key types of experimental errors. This includes erroneous assumptions in these models themselves (see note 8). Second, it provides systematic rules for generating tests with specifiably small error probabilities (e.g., so-called Uniformly Most Powerful tests).[7]

To conclude, I have argued that the ability of the Neyman-Pearson test to control low error probabilities may be seen to correspond to Peirce's self-correcting rationale of inductive methods. While, on the one hand, Neyman-Pearson methods increase the mathematical rigor and generality of Peirce's assertions about self-correcting methods, on the other, Peirce goes further in providing an experimental philosophy that can serve to justify Neyman-Pearson statistical methods in science. He does so by providing a clear answer to the question of why error probabilities matter in scientific inference: they are what permit the detection and correction of errors in making inferences from

experiments. Just as it is only by planning ahead of time that a test has a good chance of detecting errors, for Peirce "reasoning tends to correct itself, and the more so, the more wisely its plan is laid" (*CP* 5.575).[8]

Notes

This research was conducted during tenure in a National Endowment for the Humanities Fellowship for College Teachers. I gratefully acknowledge that support. I thank Nicholas Rescher for useful comments on an earlier draft.

1. See Pearson's (1955) response to Fisher, particularly p. 204.

2. In "Small Differences of Sensation" (*CP* 7.21–35), Peirce gives an interesting example of an experimental test that he himself conducted which satisfies this Neyman-Pearson requirement.

3. A "rule" for Peirce here is an assertion such as "most *A*'s are *B*'s."

4. Here Pearson is really only *asking* if this is not the real reason we consider hypothetical repetition (in contrast to a concern with a low error rate in the long run), because he wants to avoid being dogmatic. However, it is apparent from many of Pearson's examples that he values this inferential function served by considering hypothetical repetitions.

5. For a fuller argument, see Mayo 1992.

6. Except possibly in the case of observing a sample mean to estimate a population mean—sometimes referred to as quantitative induction by Peirce. The continual taking of samples yields an estimated mean that gets closer and closer to the true population mean.

7. A test with a specified Type I error is Uniformly Most Powerful if at the same time it has the smallest Type II error over the possible alternative hypotheses.

8. Peirce continues: "Nay, it not only corrects its conclusions, it even corrects its premisses." This gives further insight into Peirce's idea of self-correcting, namely, that the methods of testing can be used to check a test's own assumptions, such as that of randomness. Elsewhere Peirce discusses how randomness may be checked, once again anticipating aspects of modern sampling statistics.

References

Birnbaum, A. 1969. Concepts of Statistical Evidence. In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, ed. S. Morgenbesser, P. Suppes, and M. White, 112–43. New York: St. Martin's Press.

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

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

Levi, I. 1980. Induction as Self Correcting according to Peirce. In *Science, Belief, and Behavior: Essays in Honor of R. B. Braithwaite*, ed. D. H. Mellor, 127–48. Cambridge: Cambridge University Press.

Lindley, D. V. 1976. Bayesian Statistics. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, ed. W. L. Harper and C. A. Hooker, 2:353–63. Dordrecht: D. Reidel.

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

———. 1991. Novel Evidence and Severe Tests. *Philosophy of Science* 58:523–52.

———. 1992. Did Pearson Reject the Neyman-Pearson Philosophy of Statistical Testing? *Synthese* 90:233–62.

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

———. 1971. Foundations of Behavioristic Statistics. In *Foundations of Statistical Inference*, ed. V. P. Godambe and D. A. Sprott, 1–13. Toronto: Holt, Rinehart and Winston of Canada. Comments and Reply, 14–19.

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

———. 1966a. The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 × 2 Table.

Biometrika 34 (1947):139–67. In Pearson 1966c, pp. 169–97.

———. 1966b. The Efficiency of Statistical Tools and a Criterion for the Rejection of Outlying Observations. (Assisted by C. Chandra Sekar.) *Biometrika* 28 (1936):308–20. In Pearson 1966c, pp. 118–30.

———. 1966c. *The Selected Papers of E. S. Pearson*. Berkeley: University of California Press.

———. 1966d. Some Thoughts on Statistical Inference. *Ann. Math. Statistics* 33 (1962):394–403. In Pearson 1966c, pp. 276–83.

Rescher, N. 1978. *Peirce's Philosophy of Science*. Notre Dame, Ind.: University of Notre Dame Press.