CHAPTER ELEVEN

Why Pearson Rejected the Neyman-Pearson (Behavioristic) Philosophy and a Note on Objectivity in Statistics

> The two main attitudes held to-day towards the theory of probability both result from an attempt to define the probability number scale so that it may readily be put in gear with common processes of rational thought. For one school, the degree of confidence in a proposition, a quantity varying with the nature and extent of the evidence, provides the basic notion to which the numerical scale should be adjusted. The other school notes how in ordinary life a knowledge of the relative frequency of occurrence of a particular class of events in a series of repetitions has again and again an influence on conduct; it therefore suggests that it is through its link with relative frequency that a numerical probability measure has the most direct meaning for the human mind.
>
> —E. S. Pearson, "On Questions Raised by the Combination of Tests Based on Discontinuous Distributions," p. 228

11.1 Introduction

The two main attitudes Pearson is speaking of correspond to two views of the task of a theory of statistics: the evidential-relation or E-R view and the error probability view. We have traced the key ways in which disputes about methodological rules reflect this underlying distinction in aims. Philosophers of induction, we said, have typically embraced the first of these two views. My primary aim has not been to settle this question of aims, but rather to show how a number of disputes in philosophy of science reflect this difference in aims, and to build an account of experimental learning based on the error statistics approach. I am also concerned with showing that the error approach is at the heart of the widespread applications of statistical ideas in scientific inquiry, and that it offers a fruitful basis for a philosophy of experiment.

Despite the widespread use of error statistical methods, the official school of inference in which they are formally couched—Neyman and Pearson (NP) statistics—has been the subject of enormous controversy and criticism. From the philosophy of statistics debates of the '70s and early '80s, NP theory emerged with several black eyes, spurring on the popular new Bayesian Way. Fetzer (1981); Hacking (1965); Kyburg (1971, 1974); Levi (1980a); Rosenkrantz (1977); Seidenfeld (1979a); and Spielman (1973); as well as several statisticians have raised doubts about the appropriateness of NP theory for statistical inference in science. In a 1977 issue of *Synthese* devoted to the foundations of probability and statistics, Neyman expressed surprise at the ardor with which subjectivists (e.g., de Finetti 1972) attacked NP tests and confidence interval estimation methods:

> I feel a degree of amusement when reading an exchange between an authority in "subjectivistic statistics" and a practicing statistician, more or less to this effect:
>
> *The Authority:* "You must not use confidence intervals; they are discredited!"
>
> *Practicing Statistician:* "I use confidence intervals because they correspond exactly to certain needs of applied work." (Neyman 1977, 97)

Neyman's remarks hold true today. The subjective Bayesian is still regarded, in many philosophy of science circles, as "the authority" in statistical inference, and yet scientists from increasingly diverse fields still regard NP methods (e.g., confidence intervals) as corresponding exactly to their needs.

Howson and Urbach (1989) have attempted to renew the old efforts to cleanse science of NP methods, declaring "that the support enjoyed by classical methods of estimation among statisticians is unwarranted" (p. 198). These, along with the other NP methods, they apparently feel, should be taken to the dump heap and replaced with their brand of subjective Bayesianism. Given the new emphasis philosophers of science have placed on taking cues from actual scientific practice, this disregard if not outright condemnation of procedures that are widely and successfully used across a vast spectrum of science is curious and out of place. I think it is time to remedy the situation. Philosophers of statistics can no longer operate on the image of the philosopher issuing pronouncements on the appropriateness of the scientist's tools—not if they want to contribute to an experimental methodology that will be of relevance to science.

Much of the reason philosophers have rejected NP methods may be traced to the difference in aims just mentioned: these philosophers

seek an E-R view and NP does not give them one. To a large extent, such criticisms stem from holding to a certain philosophical image of the "logic" of statistical inference—that it should mirror deductive logic only with degrees—and not at all from finding these methods unproductive in scientific applications. In this view, a theory of statistical inference must provide a quantitative measure of evidential relationship—an E-R measure (whether a measure of support, confirmation, probability, or something else). From this perspective, NP methods will be judged inadequate for statistical inference unless NP error probabilities can be interpreted as E-R measures. Unsurprisingly, as critics show, if error probabilities (e.g., significance levels) are interpreted as E-R measures, misleading and contradictory conclusions are easy to generate. Such criticisms are not really criticisms but flagrant misinterpretations of the quantities in error statistical methods—misinterpretations repeatedly warned against in good textbooks on statistics. I have discussed criticisms based on E-R misinterpretations of error probabilities at length elsewhere (e.g., Mayo 1980, 1981, 1982, 1983, and 1985a), and I will not give them much additional consideration.

A second set of criticisms that can also be seen to follow from the E-R image of statistics is that based on assuming the likelihood principle. Since this assumption, we saw, is tantamount to assuming the irrelevance of outcomes other than the one observed, and therefore to rejecting error probabilities, these criticisms beg the question against error statistical methods. To remind us, recall the criticism of error statistical methods based on the "argument from intentions" discussed in section 10.3. If one adheres to the likelihood principle (as Bayesians do), then it does not matter whether data arose from a try and try again method or from a nonsequential experiment—the stopping rule is irrelevant. To deem stopping rules relevant—as statistical significance tests do—is, from the Bayesian point of view, tantamount to making the experimenter's intentions relevant. All the other error statistical properties are similarly found to be "incoherent" on the likelihood principle. The tables are turned completely, we saw, for an error statistician. Given an observed outcome x, the error statistician finds it essential to consider the other outcomes that could have resulted from the procedure that issued x. Ignoring aspects of the experiment that alter error probabilities (e.g., the stopping rule) violates error statistical reasoning and permits systematically misleading results.

However, we can separate out from the critical literature several legitimate questions of the epistemological basis of the NP methods: How should test results be interpreted in scientific contexts? What is so good about tests that are good or "best" on error-probability criteria? How can any of the seemingly arbitrary choices of tests and error prob-

abilities be justified? I grant that without adequate answers to these questions, the NP prescriptions can appear to license counterintuitive and unsatisfactory results.

The problem stems from the decision-theoretic framework in which NP methods are standardly couched. Although this framework has its uses, it does not adequately reflect most of the reasons that scientists find these methods correspond precisely to their needs. We need a framework that captures the nature and rationale of NP methods in scientific practice.

Happily, we already have it. In the error statistical account, formal statistical methods relate to experimental hypotheses, hypotheses framed in the experimental model of a given inquiry. Relating inferences about experimental hypotheses to severe tests of primary scientific claims is, except in special cases, a distinct step. Standard statistical ideas and tools enter into this picture of experimental inference in a number of ways, all of which are organized around the three chief models of inquiry. Their role is to (i) provide techniques of data generation and modeling along with tests for checking whether the assumptions of data models are met; (ii) provide tests and estimation methods that allow control of error probabilities; and (iii) provide canonical models of local experimental questions with associated tests and data modeling techniques.

Knowing what we want from our statistical theory, and having the elements of our framework at our disposal, it will be easy to cut through the seemingly complex arguments from philosophy of statistics. Getting NP tests to do what we want them to do, however, requires diverging from some of the key tenets that are presumed to be integral to the NP theory. The focus in this chapter is tests. The key tenets of NP testing from which we may be required to diverge are at the same time at the heart of many of the criticisms of NP theory. Accordingly, my reformulation of NP statistics will simultaneously respond to two challenges: how to answer the main criticisms of that approach, and how error statistical methods provide the needed tools for learning from error.

While it seems correct to call my approach a reinterpretation of NP statistics, I want to argue that the appropriate use of NP methods is already to be found—albeit only by hints and examples—in one of the two founders of NP statistics: Egon Pearson (as well as in most of actual practice). Egon S. Pearson (not to be confused with his father, Karl[1]), although one of the two founders of NP methods, rejected the statisti-

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

cal philosophy that ultimately became associated with NP statistics—or so I shall argue. Many contemporary criticisms of NP methods mirror Pearson's own reasons for this rejection. Extricating the view E. S. Pearson *did* hold gives a much deeper and more accurate understanding of NP principles than that which comes out in either statistics textbooks or in the presentations of critics of the NP approach. It is against these caricatures of NP methods that the criticisms of NP are largely directed. Understanding Pearsonian statistics shows how and why actual uses of NP methods generally circumvent the pitfalls without forfeiting what is central to error statistical methods: the fundamental importance of error probabilities.

11.2 NEYMAN-PEARSON THEORY OF STATISTICAL TESTS (NP TESTS)

I want to begin by putting aside for a moment the concepts of our new framework and broaching NP tests in their more formal rubric. I want to get us to consider the tests in their naked mathematical form, the better to see the latitude for their use and interpretation. The highlights of chapter 5—the examples of NP tests, the discussion of probabilistic models, and the hierarchy of models in experimental inquiry—prepare us for each of the ideas we now need. As we proceed, the connection with severity and arguments from error will emerge.

To really get down to the bare bones, the NP testing theory can be seen to define mathematical functions on random variables. The variables may take on different values corresponding to different outcomes of an experiment. Tests are functions that 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 X. As before, I confine myself to cases with only a single unknown parameter, say μ. A test is like a postal system wherein different values of X (different addresses) get sent to different values of μ (different destinations).

An example already considered several times is the Binomial experiment, the common exemplar being coin-tossing. Here the statistical variable might be the proportion of heads in n tosses, written as $\overline{X}$, and the hypotheses, assertions about the (Binomial) parameter p, the probability of heads on each toss. The test is a rule that "sends" the different observed proportions of heads to various values of the parameter p.

The standard NP test splits the possible parameter values into

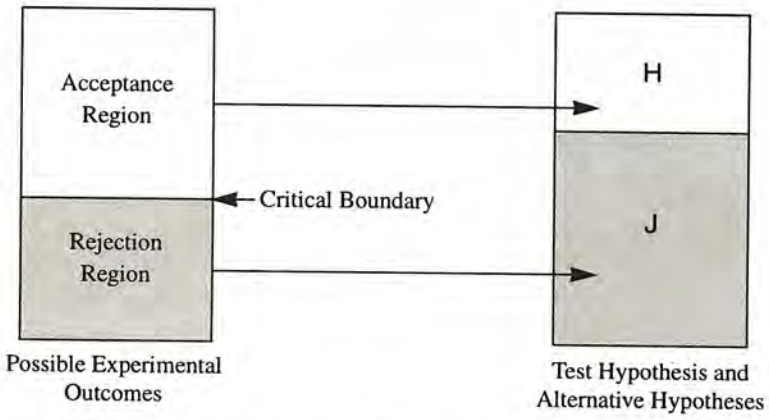

FIGURE 11.1. NP tests as mapping rules.

two—there are, so to speak, two destinations. One represents the *test hypothesis H*, the other the set of *alternative hypotheses J*. For example, *H* might assert that $p = .5$, while *J*, that $p > .5$. Hypothesis *H* here is *simple* because it consists of just one value of *p*, while *J* is *composite*. The test maps each of the possible outcomes—the experimental *sample space*—into either *H* or *J;* those mapped into *H* (i.e., into "accepting" *H*) form the *acceptance region*, those mapped into alternative *J* form the *rejection (of H) region*. This partition of the sample space is typically performed by specifying a cutoff point or *critical boundary* $\overline{X}^*$. Any outcome falling outside bound $\overline{X}^*$ falls into the rejection region.

An example would be to reject *H* whenever the observed proportion of heads, $\overline{X}$, is at least .8. The critical boundary $\overline{X}^*$ is .8. There are two ways to specify the critical boundary. The critical boundary may be given by specifying a *distance measure D* between $\overline{X}$ and *H*, and indicating "how far" $\overline{X}$ can be from *H* before slipping into the rejection (of *H*) region. Equivalently, the cutoff point may be given by specifying the significance level α, such that once that level is reached, *H* is rejected. (Recall that the larger the difference *D*, the smaller the significance level.) Leaving these acceptances and rejections uninterpreted for now, the formalism of the NP model simply describes the partitioning that results from the mapping rules as illustrated above (fig. 11.1).

NP tests focus on the probabilistic properties of these mapping rules, that is, on the probabilities with which the rule leads to one or another hypothesis, under varying assumptions about the true hypothesis. Two types of errors are considered: first, the test leads to reject

H (accept *J*) even though *H* is true (the type I error); 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, is fixed at some small number, such as .05 or .01. In other words, the test is specified to ensure that it is very improbable for an outcome to fall in the "rejection (of *H*) region" when in fact the hypothesis *H* is correct. Having fixed α, called the *size* or *significance level* of the test, NP principles seek out the test that at the same time has a small probability, 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. That is:

P(test T rejects $H \mid H$ is true) $\leq \alpha$ = probability of type I error.
P(test T accepts $H \mid J$ is true) $\leq \beta$ = probability of type II error.

When, as is quite common, alternative *J* contains more than a single value of the parameter, that is, when it is *composite,* the value of β varies according to which alternative in *J* is true. α and β are the test's formal *error probabilities.* To reemphasize, error probabilities are not probabilities of hypotheses, but the probabilities that certain experimental results occur, were one or another hypothesis true about the experimental system. Consider, for example, the probability of a type I error in testing *H* with test *T.* This is the probability of getting an experimental result that test *T* maps to "reject *H,*" when in fact *H* is true.

This leads to the cornerstone of NP 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 about the bare bones of NP tests can be summarized as follows:

> An *NP 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 percent) and β (100 percent) 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*.

Note that the size of a test is the same as the significance level of the cutoff point beyond which *H* is rejected. That is why tests with size α are often described as tests with significance level α. The relationship

between severity and size and power will be discussed explicitly in section 11.6.

11.3. The Behavioral Decision Philosophy: NP 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, "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" (Neyman 1950, 258). These are not decisions to accept or 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., 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 will 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 behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H; if $x \leq x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule . . . we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false. (Neyman and Pearson 1967b, 142)

Tests when interpreted as rules of *inductive behavior* make up a key portion of the *behavioristic (or behavioral) model* of tests. Because this model is typically associated with Neyman and Pearson theory, defects of that model are taken as defects of the theory. My position is that there are other, more satisfactory models to direct the use and interpretation of the NP methods, and that they are provided by the present

approach to experimental learning. But before getting to that it is important to do battle with a certain, not uncommon, misunderstanding.

What the Behavioral Model Does Not Say

The misunderstanding concerns the construal of "accept" and "reject" on the behavioristic model. Actually, Neyman is quite clear on what he intends. Accept *H*, Neyman says, means *to take action A* rather than *B*. Accept *H* does not mean believe *H* is true. Accept *H* does not mean act as if you knew *H* was true, in the sense of behaving in any and all of the ways you would if you knew that *H* was true. Supposing that the NP model intends this last interpretation of Accept *H*, Howson and Urbach dismiss NP theory as inappropriate for science as well as for practical action. If a scientist were to interpret accepting a statistical hypothesis in the way Howson and Urbach think NP theory intends,

> he would never bother to repeat the experiment. Moreover, he would be happy to stake his entire stock of worldly goods . . . on a wager offered at odds of, say, 10 to 1 against that hypothesis being true. Or suppose a food additive conjectured to be toxic were subjected to a trial involving 10 persons and the conjecture were rejected, then the manufacturer would be prepared to go directly into large-scale production and distribution. This interpretation of acceptance and rejection has merely to be stated to reveal its absurdity. . . . Nevertheless, despite its immense implausibility, this seems to be the way statisticians standardly interpret the notions. (Howson and Urbach, 1989, 162–63)

Their hilarious portrayal of the way they suppose "statisticians standardly interpret" the acceptance of a statistical hypothesis has no relation to any real statistician. This is not just because in reality statisticians do not strictly follow Neyman's behavioristic model, but because no such interpretation is licensed by that model. Howson and Urbach confidently assert, but on what basis I cannot imagine, that

> it is evident that the behaviour Neyman and Pearson had in mind was the acceptance and rejection of hypotheses as being true or false, that is, the adoption of the same attitude towards them as one would take if one had an unqualified belief in their truth or falsehood. (Ibid., 163)

But this is not at all evident, and Neyman and Pearson could not have been clearer in their rejection of anything like the construal that Howson and Urbach pin on the NP approach. (Nor do any of Howson and Urbach's citations offer any evidence otherwise.)

Neyman's behavioristic model literally identifies the acceptance of

H with the adoption of a decision to take some specific action A (rather than B) where A is set out at the start. One cannot choose or even articulate a test of any hypothesis until one identifies the acceptance of H with some one action A. Only then is it possible to determine the test's error probabilities, the basis upon which the choice of the test depends. The test's error probabilities may be acceptably low as regards one action while unacceptably high as regards some other. (For instance, they might be acceptably low regarding deciding to do further research on a topic, while unacceptably high regarding taking the act of publishing the results.) In the behavioral model of tests, "accept H" gets its interpretation from the specific action (pre)designated by the test in question. So the Howson-Urbach reading conflicts with the idea of fixed, predesignated error probabilities, aside from being a perversion of both Neyman's and Pearson's views.

Admittedly, from the fact that "accept H" gets its interpretation from the specific action designated by a given test, it follows that its meaning varies in different tests of H. In the behavioral model of tests, its meaning will vary according to the action identified with the result "accept H." This is just what Neyman intends.

Isaac Levi (1980a, 1984) offers perhaps the clearest depiction of the behavioristic model of NP among contemporary philosophers of statistics. He suggests "that a good approximation to [Neyman and Pearson's] intent is obtained by construing them as recommending the use of programs for using observation reports as inputs into programs designed to select acts" (Levi 1980a, 406). The idea is to have a rule, laid out ahead of time, for which action to take upon the occurrence of each possible experimental result. Such a "routine" procedure contrasts with what Levi calls a "deliberational" procedure. Where Levi and I may disagree, if we do, is on whether NP theory also admits of a nonbehavioristic (and deliberational) interpretation. Neyman himself is quite clear about his philosophy of inductive behavior, and I want to look a little at what he says.

Neyman and His Inductive Behavior

Neyman's idea of a rule of behavior is innocuous enough. Humans notice certain fairly stable patterns, Neyman begins—for example, that rain or snow storms follow the appearance of heavy clouds—and form various habits in regard to them—for example, taking cover at the sight of dark clouds. A similar kind of regularity is recognized, Neyman says, in the relative frequency with which a result occurs in repeated trials of some game of chance (real random experiments). Mathematical statistics developed as a way of providing systematic rules for how

to act with regard to this latter type of regularity. They are like rules for good habits.

Neyman offers the following, very general definition of a rule of inductive behavior:

> Let $E_1, E_2, \ldots, E_n, \ldots$ be all possible different outcomes of an experiment or of observations relating to some phenomena. Let $a_1, a_2, \ldots, a_m, \ldots$ be all the different actions contemplated in connection with these phenomena.
> *If a rule R unambiguously prescribes the selection of action for each possible outcome E_i, then it is a rule of inductive behavior.* (Neyman 1950, 10)

The statistical test, then, is a special case of a rule of behavior, one where the outcomes occur with some probability, that is, the experimental variable E follows some probability distribution. The acts, on Neyman's model, are condensed into two, a_1 and a_2. The hypotheses are assertions about the probabilities of the possible outcomes E_i—they are statistical hypotheses—and the desirability of performing the two actions depends upon which statistical hypothesis is true (Neyman 1950, 258). Correspondingly, the set of (admissible) hypotheses is split up into two, H and J, where J is regarded as not-H. The idea is that if hypothesis H (or any of the hypotheses making up region H) is true, then action A would be preferable to B, while if any of the hypotheses in J are true, action B would be preferable to A. A rule of inductive behavior determining the choice of A or B according to the experimental outcome E is a test of a statistical hypothesis.

Why does Neyman call them rules of "inductive behavior" as opposed to, say, test rules? He is led to this term because of his scruples about the term "inductive inference." Neyman begins *First Course in Probability and Statistics* as follows:

> Claims are occasionally made that mathematical statistics and the theory of probability form the basis of some mental process described as "inductive reasoning." However, in spite of substantial literature on this subject, the term "inductive reasoning" remains obscure and it is uncertain whether or not the term can be conveniently used to denote any clearly defined concept. On the other hand, as was first remarked in 1937, there seems to be room for the term "inductive behavior." This may be used to denote the adjustment of our behavior to limited amounts of observation. (P. 1)[2]

In addition to wanting to highlight the contrast with "inductive inference," Neyman was doubtless influenced by the common parlance

2. Neyman's reference is to Neyman 1967a, 250–90.

during the time NP tests were being developed. As Alan Birnbaum notes, "the 1920's and 1930's were a period of much critical concern with the meanings and possible meaningless[ness] of terms. . . . These concerns were usually pursued in terms of such doctrines as behaviorism, operationalism, or verificationism" (Birnbaum 1977, 33).

The idea of tests as rules of behavior is not all there is to the behavioristic model of tests. The other key features come in when considering how to select which of the many possible test rules to employ. In selecting such a rule one is led to consider that there are four possible situations that can result. To paraphrase Neyman 1950, p. 261:

I. Hypothesis H is true and action A is taken.
II. Hypothesis J is true (H is false) and action A is taken.
III. Hypothesis H is true and action B is taken.
IV. Hypothesis J is true and action B is taken.

These can be represented using the familiar 2 by 2 square:

	Action A	Action B
H	I	III
J	II	IV

It is assumed that A is preferable to B when H is true and that B is preferable to A if J is. As such, when a test results in situations II and III, the test errs by instructing one to take the less preferred action. The test rule is to be selected in such a way as to control the probabilities of these two types of errors. There are, however, several ways of doing this.

Neyman is led by the consideration that "with rare exceptions, the importance of the two errors is different, and this difference must be taken into consideration when selecting the appropriate test" (Neyman 1950, 261). Typically, he finds, one of the two errors is "more serious," more desirable to avoid. The behavioral model instructs one to let H—the test hypothesis—be the one whose erroneous rejection is considered the more serious. (Situation III is worse than situation II.) This error—the one that comes first in importance—is to be made the type I error of the test. The test is selected to fix the probability of the type I error at some low value and then choose the test that does best (or at least reasonably well) as regards the probability of a type II error.

The paradigm example that seems to fit the behavioristic model is acceptance sampling in industrial quality control. Here a sample from some batch of products is observed in order to decide whether or not to reject the batch as containing too many defectives, say, for shipping.

This is a paradigmatic case in which the importance of errors reflects economic values, and the differential weighing of the errors reflects the losses judged affordable. The values can also be ethical, as in one of Neyman's main illustrative examples.

Testing the Toxicity of Drugs: Neyman

In manufacturing drugs, impurities occasionally enter that are sufficiently toxic that minute quantities that escape ordinary chemical analysis can be dangerous. Prior to putting a newly manufactured lot on the market, it is tested. Small doses are injected into experimental animals and the effect recorded. Let X, the variable recorded, be the number of deaths among the n animals injected with a specified dose of the drug. The experiment is modeled as observing this random variable X. The probability of the different possible number X of deaths depends on how toxic the drug is: all or most animals die if the drug is toxic. The different values of X, Neyman supposes, may lead to one of two possible courses of action: (a_1) put the lot of drug on the market, and (a_2) return the lot to the manufacturer:

> The two kinds of error connected with actions a_1 and a_2 are very different. . . . First consider the case where action a_1 is taken when the appropriate action is a_2. This means that the drug is dangerously toxic but declared harmless through the unavoidable inaccuracies of the experiment. . . . Error of this kind may cause death to the patients treated with the drug. Actual cases of this kind are on record. (Neyman 1950, 263)

Neyman contrasts this with the error of taking action a_2—returning the lot—when in fact a_1 is appropriate. Although the consequences of this error are unpleasant, and may result in financial losses to the manufacturer and an increased price of the drug, "the occasional rejection of a perfectly safe drug is clearly much less undesirable than even an infrequent death of a patient" (ibid.). So this type of error is less important than the first, and would be identified as the type II error.

Neyman's test, then, is a rule of inductive behavior with two hypotheses, two corresponding actions, and two associated errors, one (typically) more important than the other; and the basis for selecting among tests is the goal of controlling the probabilities that they would lead to these errors. In Neyman's view, "in many cases the relative importance of the errors is a subjective matter" and "lies outside of the theory of statistics" (Neyman 1950, 263). Such remarks have led to misunderstandings (see section 11.6). To understand Neyman's attitude, I suggest we think back to his view regarding the use of statistical

models generally (discussed in chapter 5). His attitude seems to be this: here is a formal statistical technique that seems to reflect certain features of a standard testing context. It is up to you to assign acts to the two hypotheses, to ascertain which of the two errors is more important to avoid in your testing context (making that one the type I error) and to determine how often such an error seems acceptable (which will direct you to fix the α level of your test). That is "subjective". The NP theory can then use its machinery to find the test that at the same time minimizes the probability of a type II error (β). Once the rule is selected (and assuming the assumptions are approximately met), hypothesis testing is on automatic pilot—on the behavioral model. Applying the test just means following the rule. The experimental outcome is observed, and the test tells you whether to take one of two actions, *A* or *B*, according to whether or not the outcome falls in the rejection region of hypothesis *H*. There is no hemming and hawing, no agonizing over the particular case. Your long-run low error rate needs are guaranteed, and they are guaranteed objectively.

All this is fine and dandy, say critics, if your actual needs correspond to the kind of decision-making context envisioned in the behavioristic model; but scientific inquiry does not seem to be such a context. The issue is not just whether science involves decision making or whether inference can be seen as a kind of decision. Many who are happy to regard all of science as decision making—the typical Bayesian decision maker—reject the NP theory, not because of its development along decision-theoretic lines, but because it does not go far enough in its decision-theoretic leanings. (A full decision theory would involve not only the losses captured in error probabilities but explicit loss functions, prior probabilities, and all the rest of the "full dress" Bayesian approach.) The issue now, raised by both Bayesian and non-Bayesian critics of the behavioristic approach, concerns the appropriateness of the particular kinds of decision strategies depicted in the behavioristic model. Letting a decision to accept or reject a hypothesis turn on whether data reaches a cutoff point just seems too, well, too automatic. Many statisticians allege that no one, not even Neyman, ever tests a scientific claim along the strict behavioristic line.

I agree with them. My position all along is that the NP account admits of a nonbehavioral construal that is more satisfactory and more accurately reflects how NP methods are used in experimental learning. By and large, however, NP tests are still formulated along the lines of the behavioristic model, with the probability of a type I error generally set at the conventional levels of .01 or .05. Why are NP methods so productively used in science despite their "rule of behavior" formulation? How, paraphrasing Neyman, do they manage to correspond pre-

cisely to the needs of applied research? There seem to be two main reasons: First, many scientific tasks fit the "assembly line" behavioral-decision model. At many junctures in the links between experimental and data models there is a need for standardized approaches to data analysis that allow us to get going with few assumptions, enable results to be communicated uniformly, and help ensure that we will not too often err in declaring "normal puzzles" solved or not. Second, the behavioral decision approach provides canonical models for nonbehavioral and non-decision-theoretic uses. The behavioral concepts simply serve to characterize the key features of NP tools, and *these features* are what enable them to perform the nonbehavioral tasks to which tests are generally put in science.

Although I take the second reason to be weightier, as well as being the more interesting for our purposes, the first reason should not be disregarded altogether. There are uses for statistics in science for which the behavioristic construal is apt and for which NP as a theory for routine decision making has made for real progress. One type of example is discussed by the statistician Irwin Bross.

Controlling the Noise in Communications Networks

The context Bross (1971) discusses concerns decisions to report a given message, say, that a drug is effective or, more generally, that an effect "is real," not spurious. Bross's particular focus is on analgesics. The act of reporting that a drug is effective is not tantamount to taking any and all acts that would be licensed were it known to be effective, a point we have already made. (The act of reporting is distinct from subsequent possible actions, say, for physicians to use the drug or to buy stock in the drug company.) But a decision to report it as effective may have repercussions for subsequent decisions, and tools for routine error control may be called for.

An NP test may be used as a routine for declaring an analgesic effective in the following manner. It may stipulate: report a drug effective only if an observed difference in effect rates is statistically significant at, say, the 1 percent level. Following Bross, the various sources of error that can creep into scientific reports may be seen as sources of noise in a scientific communications network. Noise from random sources—which is inevitable in experimental research—is often called sampling variation.[3] The adoption of a fixed critical level or size, say 1 percent, is useful in "controlling the noise in communication networks." According to Bross, prior to the advent of controlled clinical

3. Noises from nonrandom sources are sometimes called biases or extraneous variables.

trials the noise level in analgesic testing was high enough to impede progress seriously (Bross 1971, 503–4).

To illustrate, Bross describes an uncontrolled drug testing network where researchers report favorably on all new drugs tested, noting that "some years ago this would not have been an entirely unrealistic model for certain networks" (p. 504). If four out of five of these drugs are actually ineffective, being no more effective than a standard agent, 80 percent of the favorable reports would be false. With such a high level of noise, no reliance could be placed on the reports. If, on the other hand, each member of the network reports favorably only on those drugs that pass a test with critical level or size of, say, 5 percent, then the proportion of false positive reports is kept low, at 5 percent.

Thus the use of statistical significance tests as accept-reject routines for a thumbs up or down approach on analgesics helped to control the noise in the scientific communication network. True, it would not do to apply such a behavioristic construal of tests to deciding to accept or reject substantive scientific hypotheses directly. Nevertheless, the hierarchy of models in an experimental inquiry may also be seen as a "communication network," and it is plainly desirable to have tools for controlling errors at numerous points in this network of models. Controlling the errors at various segments of the inquiry is what enables the overall reliability and severity to be achieved. One could well imagine, for example, how Jean Perrin might have used a routine test to give a "yes or no" pronouncement to whether a given grain, after undergoing his technique of fractional centrifuging, was sufficiently uniform to be included in the next stage of the Brownian motion analysis (chapter 7). One act would be including the grain into the analysis, a second would be to subject the grain to further centrifuging. Assurance that he would rarely include insufficiently uniform grains as well as rarely carry out unnecessary centrifuging was precisely what Perrin sought.

These points go toward illustrating what I gave as the first reason that the behavioral model of tests has a serviceable role in research, namely, that there are scientific tasks that fit the behavioristic model. Even these uses, however, depend upon designing, interpreting, and combining several tests in a manner that is decidedly *not* automatic. The second and more important reason that NP tests supply needed tools for research is that their methods provide standard or canonical models for nonbehavioral and non-decision-theoretic uses. Undoubtedly, many of the behavioral concepts with which Neyman chose to characterize the key features of NP tests would not have been chosen by Pearson. But these concepts succeed in characterizing the features

of the tests well enough, and these features are what enable tests to perform the nonbehavioral tasks to which they are generally put in science. In these nonbehavioristic contexts, tests license not acts, but arguments or inferences as to what is learned from particular experimental results. The arguments are arguments from error.

This, I propose, is why behavioral models of tests provide serviceable canonical models for nonbehavioral tasks. The tests can and should be seen as tools whose distinctive properties enable them to be used to ask a variety of standard questions about errors—quite generally construed. The result of a single statistical test does not license a substantive scientific inference. Instead, each such test, or set of tests, teaches the answer to a specific question, and error control at local points is the key to arriving at substantive severity arguments.

11.4 PEARSON REJECTS THE NEYMAN-PEARSON (BEHAVIORISTIC) PHILOSOPHY

Alan Birnbaum (1969, 1977) argued that NP admits of two types of interpretations: in one, Neyman's behavioral decision view, we saw that the test result is literally a decision *to act* in a certain way; in 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.[4] While I do not embrace the particular evidential interpretation Birnbaum favored, I think he was quite right that in situations of scientific research the behavioral interpretation of tests is

4. Birnbaum called the concept underlying this evidential interpretation of NP the *confidence concept*, which he formulated (1977, 24) 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.

Birnbaum argued that scientific applications of NP tests made intuitive use of something like the confidence concept. Birnbaum's approach, incomplete at the time of his death, sought to make explicit the correspondence between an NP result and a statement about the 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 .01 and .2, respectively, as very strong statistical evidence for H as against J. A main shortcoming, as I see it, is that it interprets a test output—say, reject H—from two tests with the same α, β as finding equally strong evidence for J. Depending upon the particular outcome and the test's sample size, the two rejections may constitute very unequal tests of J—something I take up in later sections. Birnbaum's rules do not seem to reflect such differences. Further criticism along these lines occurs in Pratt 1977. I discuss more generally attempts at "evidential" interpretations of NP methods in Mayo 1985a.

intended to apply "in a way which is heuristic or hypothetical, serving to explain the inevitably abstract theoretical meanings associated with the error probabilities [and] formal 'decisions' such as 'reject H''" (Birnbaum 1977, 32–33). The behavioristic formulation of tests, Birnbaum proposed, should simply be seen as a way of articulating the new statistical ideas of the NP approach. That the behavioral construal of tests is still with us, I suggest, testifies that they still serve the kind of heuristic function that Birnbaum had in mind.

Birnbaum found clues of these nonbehavioral intuitions in the writings of Pearson. One particularly interesting document that Birnbaum (1977, 33) supplies includes an unpublished remark by Pearson in 1974:

> 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)

Pearson never articulates just what evidential interpretation he supports, and I do not think that Birnbaum's evidential model, so far as he worked it out (in which NP results are reinterpreted in terms of strong or weak evidence for hypotheses), is indicated in Pearson's "signs of evidentiality." Nevertheless, I endorse Birnbaum's proposal that the behavioral model of NP tests be regarded as a device to communicate what the tests could be used for, while requiring reinterpretation in scientific contexts. This, I believe, was also Pearson's view, and that is why I say Pearson rejects what have come to be identified as the key tenets of the NP behavioral philosophy. What Pearson rejects is the philosophy associated with Neyman's inductive-behavior model.

The Rationale of Tests according to the NP Behavioristic Philosophy

Because NP theory developed mathematically in a decision-theoretic framework (along with the work of Abraham Wald), the statistical philosophy generally associated with these tools is Neyman's behavioral decision one. Often it is referred to as the Neyman-Pearson-Wald (NPW) approach.[5] We can identify two closely connected aspects

5. Even that arch opponent of Neyman, Bruno de Finetti, held that the expression "inductive behavior . . . that was for Neyman simply a slogan underlining and explaining the difference between his own, the Bayesian and the Fisherian formulations" became, with Wald's work, "something much more substantial" (de Finetti 1972, 176). He called this "the involuntarily destructive aspect of Wald's work" (ibid.).

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. While these features, taken strictly, give a caricature of tests, even as Neyman intended them, they are at the heart of the philosophical criticisms of NP to which we need to respond.

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 (apparently) solely their ability to guarantee low long-run errors in some sequence of applications. This is not a final measure of the degree of support or probability acquired by hypotheses—it is not an E-R measure. For example, to reject H with a test having a low probability of erroneous rejections does not say that the *specific* rejection has a low probability of being in error, but only that it arises from a testing *procedure* that has a low probability of leading to erroneous rejections. Likewise with confidence levels attached to particular interval estimates. Critics of NP theory deny that low error rates in the long run are relevant to justifying any particular inference.

Tests as decision "routines" with prespecified error properties. This feature is associated with two main criticisms. First, there is the fact that the NP decision model does not give an interpretation customized to the specific result. A test result either is or is not in the prespecified 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 prespecified *size*) is identical for any such rejection.[6] Second, there is the problem of how to interpret test results. Deciding to accept or reject hypotheses, construed as deciding how to act, does not seem to offer the kind of evidential appraisal needed for scientific inference.

A Dialogue between Pearson and Fisher

These features are not only the source of contemporary criticisms of NP theory. They lie at the heart of R. A. Fisher's original attack on

6. The point here is that to do no more than report the error probabilities, while condoned by the strict NP decision model, is not sufficient to discriminate between these two results—one of the sources of the criticisms of NP tests. *Other* uses of error probabilities, however, can make this discrimination along the lines I discuss in sections 11.6 and 11.7.

Neyman and Pearson's reworking of (what Fisher regarded as "his") significance tests. In his forceful 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, 70)

Fisher makes a similar comparison 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.)

The allegation is essentially the one cited earlier: NP methods seem suitable for industrial acceptance sampling, but not for drawing inferences in science.

Pearson, however, responds to Fisher's attacks—something critics seem to have overlooked. Perhaps this is because it occurs in an obscure, very short (but fascinating) paper, "Statistical Concepts in Their Relation to Reality" (Pearson 1955), that is not included in *The Selected Papers of E. S. Pearson.*

Pearson Responds to Fisher

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 nondeliberational, routine use of tests. These two features are regarded as so integral to the NP model that I think it is fair to say that Pearson rejected the NP philosophy (but not NP methods).[7] 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 NP tests. These are the "signs of evidentiality" to which Pearson alluded.

Pearson's Original Heresy

Let us begin with Pearson's (1955) response to Fisher's main criticism—that the NP model turns tests into a pragmatic, five-year-plan type of a process. Pearson insists that

7. Perhaps it is clearest to say that what Pearson rejected was the Neyman-Pearson-Wald (NPW) model of NP methods. See also Note 5.

> there 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, 204; emphasis added)

Interestingly, Fisher directs his attacks at *Neyman's* behavioral approach, leaving Pearson out of it.[8] Nevertheless, Pearson protests here that the "original heresy" was really his!

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 (from Fisher) in insisting that tests explicitly take into account alternative hypotheses, in contrast with Fisherian significance tests, which did not. With just the single hypothesis (the null hypothesis) of Fisherian tests, the result is either reject or fail to reject according to the significance level of the result. However, just the one hypothesis and its attended significance level left too much latitude in specifying the test, rendering the result 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.

In responding to Fisher, Pearson is not merely arguing that NP methods *can* be interpreted in a manner other than a pragmatic behavioral-decision one, he is claiming that their original formulation (admittedly "heretical" in the above sense) was not even intended to capture decision-theoretic aims. Those aims came later, and were not his:

> 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., 204)

Having sketched for Fisher this marvelous image of his sitting on a gate (my own sketch being the frontispiece), Pearson goes on to explain that his thoughts had not at all to do with speeding up production or saving money. Rather, Pearson continues,

> *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.*

8. George Barnard, in a private communication, revealed the part he played in Fisher's reception of NP theory. It was Barnard who alerted Fisher to the consequences of proceeding within the behavioristic model of tests favored by Neyman. At the same time, Barnard told Fisher that Neyman's model was to be distinguished from Pearson's philosophy. Barnard 1985 provides an excellent discussion of historical developments in statistics, as well as comments from a number of statisticians.

> 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.* (Pearson 1955, 204–5; emphasis added)

So it was only after the main NP theory had taken shape that a "remarkable parallelism" with acceptance sampling problems was discovered. And while the NP methods clearly benefited from the mathematical rigor of the newly developed work in decision theory, the original application, as Pearson saw it, was to learning from data in science.

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 again 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., 206; emphasis added)

Thus, says Pearson, the NP framework, with its consideration of alternative hypotheses, grew out of an attempt to provide tests then in use with an epistemological rationale—one based on their function as learning tools. In this role, the test's output was not supposed to be identified with the final acceptance or rejection of a scientific hypothesis. Instead, the test teaches about a specific aspect of the process that produced the data. A suitable reformulation of NP tests, I believe, grows directly out of the distinct roles that statistical tests play in filling out and linking models in an experimental inquiry. Although Pearson did not himself propose such a reformulation, Pearson clearly distances the original learning function of NP methods from the later behavioral-decision construal to which Fisher is objecting. He declares in the last line of this paper that

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

Pearson Rejects the Long-Run Rationale

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

> How far then, can one go in giving precision to a philosophy of statistical inference? (Pearson 1966a, 172)

He considers the rationale that might be given to NP 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., 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 judgement which we control at a low figure? (Ibid., 173; emphasis added)

Although Pearson leaves this tantalizing question unanswered, claiming, "On this I should not care to dogmatize," it is evident from his treatment of type *B* cases that, for Pearson, "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 what he calls the "commonsense" objection to long-run justifications—precisely the objection lodged by contemporary critics of NP theory:

> 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., 171)

Seeing how the scientist makes his contribution leads to substantiating my second claim, that Pearson rejects the routine use and interpretation of NP tests associated with the behavioral model. For the scientist's "contribution to a round-table discussion" turns on the thoughtful use of error probabilities to unearth causal knowledge—something not reducible to routine.

Nonroutine Uses of Tests: An Example of Type B

Weaving together strands found throughout Pearson's work, one can craft a picture of statistical tests much like the one I would promote, namely, as tools for learning about causal processes by enabling a piecemeal series of standard questions (about errors) to be posed and reliably answered. In the opening of a 1933 paper (jointly written with S. S. Wilks) Pearson writes:

> 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 1966, 81; emphasis added)

An example that Pearson often employs as a case of type *B*, where no repetition is intended, is the following:[9]

> *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 1966a, 171)

Pearson's interest in this naval shell example stems from his own work on the statistical assessment of army weapons in World War II and after. The experimental 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 $\frac{11}{24}$ (i.e., $\frac{10}{12} - \frac{3}{8}$). So

9. Pearson follows this naval shell example through a number of papers.

we have a standard case of a difference in proportions similar to our birth-control pill example in chapters 5 and 6. (In both cases, the null hypothesis predicts a zero difference.) Statistical tests aid the scientist's contribution here by answering a question under (*b*) about the causal 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., 171)

Notably, Pearson does not simply report whether 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" (p. 192) if hypothesis *H* were true and there was no difference in piercing qualities. This is, we know, the *significance level* of the observed difference—a measure that reflects the actual result observed.

Although testing the "no difference" hypothesis is standard, there is not just one plausible way to test it. More than one way has been proposed to describe the data and define a distance between data and hypotheses. This matter is the basis of a historical debate between Pearson and others, which I leave to one side. Although Pearson takes a position in this debate (arguing in favor of the test that he regards as more nearly describing the experimental situation), he does not feel that a single best test needs to be found. Pearson is not perturbed by the existence of this latitude in choosing tests, he does not see it as presenting a problem. It would only present a problem, he thinks, to one who regards tests as giving automatic routines; but, in striking contrast with the routine decision model, Pearson held that little turns on which of the various plausible tests one employs. Treating the (difference between two proportions) case in one way, Pearson obtains an observed significance level of .052; treating it differently (along Barnard's lines), he gets .025 as the (upper) significance level.[10] In an auto-

10. The first treatment falls under what Pearson calls Problem I (Barnard's "2x2 independence trial"). Here the question is restricted to just the 20 shells observed, the total number of failures being fixed at the observed one, 7. The test asks whether the observed difference is due to a random partition of the 20 individual shells, of whom 7 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 to 12 and 0 to 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 "2x2

matic routine use of tests this can make a substantial difference. Pearson rejects this use of tests.

> The result of either approach would raise considerable doubts as to whether the performance of the [second] type of shell was as good as that of the [first]. (Ibid., 192)[11]

In either case, the data indicate *J:* the first type of shell is better than the second, because in either case *J* passes a severe test (although one is more severe than the other). (Severity for passing *J* here is 1 minus the significance level.)

Pearson holds that in important cases the difference in error probabilities, depending upon which of these tests is chosen, makes no real difference to substantive judgments in interpreting the results:

> 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., 192; emphasis added)

So, faced with this type of investigation, no responsible statistician would be a strict follower of the behavioristic model of NP tests.

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

> If then a statistician thoughtlessly decides, whatever be the test, to reject an hypothesis when $P \leq .01$, 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 1967c, 18)

---

comparative trial"). For the naval shell example, Pearson regards the former treatment, though preferred by Barnard, as more artificial than the latter. Which of several ways to treat the 2x2 case had been much debated by Barnard and Fisher at that time. Pearson's position is that the appropriate sample space "is defined by the nature of the random process actually used in the collection of the data," which in turn directs the appropriate choice of test (Pearson 1966a, 190). But Pearson does not think there is a need for a rigid choice from among several plausible tests.

11. Pearson's conclusion inadvertently switches the observation to 2 of 12 and 5 of 8 *successful* perforations, where originally they had been failures. I have stated his conclusion to be consistent with the original results reported in this example.

(*P* here is equal to the significance level.) In other words, if the decision model of NP 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 .06 while the fixed level for rejection is .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 .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., 27)

Most significant in this joint contribution is the declaration that

> if 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., 56–57)

This introduces a criterion distinct from low error rates, namely, the *relevance* of the information. In addition, clues emerge for connecting tests (used nonroutinely) to learning about causes by probing key errors:

> The 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., 58)

This sentiment is clear enough: we should not infer some alternative to a hypothesis *H* until other alternative explanations for the discordancy with *H* have been ruled out. The surprise is only that such nonbehavioral talk should occur in a joint paper. Its very title—"On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference"—is at odds with Neyman's philosophy, which concerned behavior and not inference. A curious note by Neyman tucked at the end of this paper may explain its Pearsonian flavor.

> 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 1967c, 66; signed by J. Neyman)

This "joint" paper, it appears, was largely a contribution of Pearson's.

11.5 A PEARSONIAN PHILOSOPHY OF EXPERIMENTAL LEARNING

I want now to turn to Pearson's discussion of the steps involved in the original construction of NP tests (of *H:* no difference). His discussion underscores the key difference between the NP error statistical (or "sampling") framework and approaches based on the likelihood principle. The previous chapters have amply illustrated the enormous consequences that this difference makes to an account of scientific testing. This background should let us quickly get to the heart of why different choices were made in the mathematical development of NP error statistics. The choices stem not only from a concern for controlling a test's error probabilities, but also from a concern for ensuring that a test is based on a plausible distance measure (between data and hypotheses). By recognizing these twin concerns, we can answer a number of criticisms of NP tests.

Three Steps in the Original Construction of NP Tests

After setting up the *test* (or null) *hypothesis,* and the *alternative hypotheses* against which "we wish the test to have maximum discriminating power" (Pearson 1966a, 173), Pearson defines three steps in specifying tests:

> *Step* 1. We must specify [the *sample space*[12]] 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,* on the information available, to reject the hypothesis tested in favour of alternatives which differ from it by increasing amounts. (Pearson 1966a, 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

12. Here Pearson calls it the "experimental probability set."

become more probable under the assumption that some alternative J is true than under the assumption that H is true. This suggests that one plausible measure of inclination is the likelihood of H—the probability of a result e given H. We are "more inclined" toward J as against H to the extent that J is more likely than H given e.

NP theory requires a third step—ascertaining the error probability associated with each measure of inclination (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. (Ibid.)[13]

For example, step 2 might give us the likelihood or the ratio of likelihoods of hypotheses given evidence, that is, 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, that is, an error probability. We are already familiar with this kind of calculation from calculating severity.

Pearson explains that in the original test model step 2 (using likelihood ratios) did precede step 3, and that is why he numbers them this way. Only later did formulations of the NP model begin by first fixing the error value for step 3 and then determining the associated critical bounds for the rejection region. This change came about with advances in the mathematical streamlining of the tests. 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., 173)

However, if the rationale is *solely* error probabilities in the long run, the need to *first* deliberate over an appropriate choice of measuring distance at step 2 drops out. That is why it is dropped in the standard behavioral model of NP tests. In the behavioral model, 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) NP test determines the rejection region. In other words, the result of step 3 automatically accomplishes step 2. From step 3 we can calculate how the test, selected for its error probabilities, divides the possible out-

13. 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.

comes. Yet this is different from having first deliberated at step 2 about 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, despite having adequate error probabilities, might have an inadequate distance measure. Such a test 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. The test may even be irrelevant to the hypothesis of interest. The reason that critics can construct counterintuitive tests that appear to be licensed by NP methods, for example, certain mixed tests,[14] is that tests are couched in the behavioral framework from which the task Pearson intended for step 2 is absent.[15]

Likelihood Principle versus Error Probability Principles, Again

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? This is precisely what many critics of NP have asked. It was essentially Hacking's (1965) point during his "likelihood" period. As briefly noted in previous chapters, Hacking's likelihood account held that the *likelihood ratio* (of H against alternative J) provides an appropriate measure of support for H against J.[16] In such a likelihood view, the tests *should* just report the measure of support or inclination (at step 2) given the data. For Bayesians also, the relevant evidence contributed by the data is fully contained in the likelihood ratio (or the Bayesian ratio of sup-

14. In a mixed test certain outcomes instruct one to apply a given 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 NP theory (where the lower the error probabilities the better the test) can inveigh against them. This is not the case for one who rejects the behavioral model of NP tests 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 presents a related argument against mixed tests, except that he feels it is necessary to appeal to propensity notions, whereas I appeal only to frequentist ones. See, for example, Giere 1976.

15. 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. See also note 28.

16. Hacking later rejected this approach (e.g., Hacking 1972). Although he never clearly came out in favor of NP methods, in 1980 he reversed himself (Hacking 1980) on several of his earlier criticisms of NP methods.

port)—the thrust of the likelihood principle (LP).[17] We discussed the LP at length in chapter 10. To remind us, NP theory violates the LP because a hypothesis may receive the same likelihood on two pieces of data and yet "say different things" about what inference is warranted—at least to the error statistician's ears. To pick up on this difference requires considering not only the outcomes that did occur, but also the outcomes that might have occurred; and, as we saw, the Bayesian (or conditionalist) recoils from such considerations.

The debate in the philosophy of statistics literature often does little more than register the incompatibility between the NP approach on the one hand and the likelihood and Bayesian approaches on the other. Each side has a store of examples in which the other appears to endorse a counterintuitive inference. From the perspective of the aims of ampliative inquiry, I have been arguing, we can go further: control of error probabilities has a valid epistemological rationale—it is at the heart of experimental learning. The main lines of my argument may be found in Pearson. Here is where Pearson's rejection of the long-run rationale of error probabilities and his nonroutine use of tests come together with the Pearsonian logic of test construction.

Likelihoods Alone (Step 2) Are Insufficient for Pearsonian Reasoning

Pearson explains why he and Neyman held it essential to add the error probability calculations of step 3 to the "measures of inclination" at step 2. The concern was *not* pragmatic, with low error rates (in the long run of business), but with learning from experiments. Reflecting on this question (in "Some Thoughts on Statistical Inference"), Pearson tells of their "dissatisfaction with the logical basis—or lack of it—which seemed to underlie the choice and construction of statistical tests" at the time. He and Neyman, Pearson explains, "were seeking how to bring probability theory into gear with the way we think as rational human beings" (Pearson 1966e, 277).

17. The likelihood principle, we saw in chapter 10, falls out directly from Bayes's theorem. Birnbaum is responsible for showing, to the surprise of many, that it follows from two other principles, called sufficiency and conditionality (together, or conditionality by itself). For an excellent discussion of these principles see Birnbaum 1969. Birnbaum's result—while greeted with dismay by many non-Bayesians (including Birnbaum himself) who balked at the likelihood principle but thought sufficiency and conditionality intuitively plausible—was welcomed by Bayesians, who (correctly) saw in it a new corridor leading to a key Bayesian tenet. A third way would be to steer a path between the likelihood principle and advocating any principle that decreases error probabilities, thereby keeping certain aspects of sufficiency and conditionality *when and to the extent that they are warranted.*

> *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., 277–78; emphasis added)

I have the impression that Pearson is correct. The main focus of philosophical discussions is on what rival statistical accounts tell one to do once "data are thrown at the statistician and he is asked to draw a conclusion"; for example, to accept or reject for an NP test or compute a posterior probability for a Bayesian.

Why are error probabilities so important once the "preliminary planning and subsequent interpretation" are closely linked? First, if one of the roles of a theory of statistics is to teach how to carry out an inquiry, then some such before-trial rules are needed. By considering ahead of time a test's probabilities of detecting discrepancies of interest, one can avoid carrying out a study with little or no chance of teaching what one wants to learn; for example, one can determine ahead of time how large a sample would be needed for a given test to have a reasonably high chance (power) of rejecting H when in fact some alternative J is true. Few dispute this (before-trial) function of error probabilities.

But there is a second connection between error probabilities and preliminary planning, and this explains their relevance even after the data are in hand. It is based on the supposition that in order to correctly interpret the bearing of data on hypotheses one must know the procedure by which the data got there; and it is based on the idea that a procedure's error probabilities provide this information. The second role for error probabilities, then, is one of interpreting experimental results *after* the trial. It is on this "after-trial" function that I want to focus; for it is this that is denied by non-error-statistical approaches (those accepting the LP).[18] The Bayesians, paraphrasing LeCam's remark (chapter 10), have the magic that allows them to draw inferences

18. Some (e.g., Hacking 1965) have suggested that error probabilities, while acceptable for before-trial planning, should be replaced with other measures (e.g.,

from whatever aspects of data they happen to notice. NP statisticians do not.

Throughout this book I have identified several (after-trial) uses of error probabilities, but they may all be traced to the fact that *error probabilities are properties of the procedure that generated the experimental result.*[19] This permits error probability information to be used as a key by which available data open up answers to questions about the process that produced them. Error probability information informs about whether given claims are or are not mistaken descriptions of some aspect of the data generating procedure. It teaches us how typical given results would be under varying hypotheses about the experimental process.

We know how easy it is to be misled if we look only at how well data fit hypotheses, ignoring stopping rules, use-constructions, and other features that alter error probabilities. That is why fitting, even being the best-fitting hypothesis, is not enough. Step 2 assesses the fit, step 3 is needed to interpret its import. In a joint paper, Pearson and Neyman (1967) explain that

> if we accept the criterion suggested by the method of likelihood it is still necessary to determine its sampling distribution in order to control the error involved in rejecting a true hypothesis, because a knowledge of L [the likelihood ratio] alone is not adequate to insure control of this error. (P. 106; I substitute L for their λ)

Let L be the ratio of the likelihood of H and an alternative hypothesis J on given data x. That is,

$$L = \frac{P(x \mid H)}{P(x \mid J)}$$

(where in the case of composite hypotheses we take the maximum value of the likelihood). Suppose that L is small, say, .01, meaning H has a much smaller likelihood than J does. We cannot say that because

---

likelihoods) after the trial. Pearson took up and rejected this proposal, raised by Barnard in 1950, reasoning that

> if the planning is based on the consequences that will result from following a rule of statistical procedure, e.g., 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 1966c, 228).

19. It may be objected that there are different ways of modeling the procedure. That is correct but causes no difficulty for the after-trial uses of error probabilities. Indeed, using different models is often a useful way of asking distinct but interrelated questions of the data.

L is a small value, "we should be justified in rejecting the hypothesis" H, because

> in order to fix a limit between "small" and "large" values of L we must know how often such values appear when we deal with a true hypothesis. That is to say we must have knowledge of . . . the chance of obtaining [L as small or smaller than .01] in the case where the hypothesis tested [H] is true. (Ibid., 106)

Accordingly, without step 3 one cannot determine the test's severity in passing J, and without this we cannot determine if there really is any warranted evidence against H.

The position I want to mark out even more strongly and more starkly than Pearson does is that the interest in a test's error probabilities (e.g., the probability of it passing hypotheses erroneously) lies not in the goal of ensuring a good track record over the long haul, but in the goal of learning from the experimental data in front of us. Comparisons of likelihoods or other magnitudes of fit can measure the observed difference between the data and some hypothesis, but I cannot tell if it *should* count as big or small without knowledge of error probabilities. In one particularly apt passage, Pearson explains that error probability considerations are valuable because

> *[they help] us to assess the extent of purely chance fluctuations that are possible.* It may be assumed that in a matter of importance we should never be content with a single experiment applied to twenty individuals; but *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. (Pearson 1966a, 176–77; emphasis added)

We saw how in certain cases of use-constructing hypotheses to fit data (e.g., gellerization cases), as well as in cases with optional stopping, the chance of mistaken conclusions may be very high. This error-probability information showed us how easily chance fluctuations could be responsible for a large extent of the results.

Let us go back to the case of Pearson's naval shell. The (after-trial) question asked was "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"? (Pearson 1966a, 171). It is asked by testing hypothesis H:

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

(Alternative J would assert that it is due to a systematic discrepancy in the processes, with respect to successfully piercing the plate.) The difference statistic D is the difference between the proportions of suc-

cessful perforations of the plate from the two types of shell. Using the experimental (or sampling) distribution of D he can calculate the statistical significance of a given observed difference D_{obs}:

The *statistical significance* of D_{obs} = P(a difference as great as $D_{obs} \mid H$).

He found D_{obs} to be improbably far from what would be expected were H correct. (The difference falls in the rejection region of a test of approximately .05 size.) Even if no repetitions are planned, this analysis is informative as to the origin of *this* difference. There are many ways of expressing this information.

One, paraphrasing Pearson, is that the observed difference (in piercing ability) is not the sort easily accounted for by inevitable variability in the shells and measurement procedures. The observed difference, rather, is indicative of the existence of a genuine (positive) difference in piercing ability. Were the two shells about as good, it is very probable that we would *not* have observed so large a difference—the severity in passing J is high. Finding the data indicative of hypothesis J, even with larger sample sizes than in this simple illustrative example, is just a first step. For simplicity, suppose that hypothesis J includes positive discrepancies in piercing rates between the first and second types of shell. One may also want to know which of the particular discrepancies are indicated by outcome D_{obs}. This further information may be obtained from the same experimental distribution, but the hypothesis to the right of the given bar would now be a member of J. We can thus learn how large a discrepancy in piercing rates would be needed to generate differences as large as D_{obs} fairly frequently. This calls for a custom-tailoring of the interpretation of test results to reflect the particular outcome reached. In the next section I shall consider two basic rules for interpreting test results that take into account the particular outcome observed. While they go beyond the usual NP test calculations, they fall out directly from the arguments based on severity calculations considered earlier.

11.6 TWO ERROR STATISTICAL RULES TO GUIDE THE SPECIFICATION AND INTERPRETATION OF NP TESTS

Before proceeding with our next task, let me remind the reader that it pertains to just one piece, albeit a central one, of the series of tasks to which statistical considerations are put in the present account. In this piece, which is often regarded as statistical inference proper, statistical methods (tests and estimations) link data to experimental hypotheses, hypotheses framed in the experimental model of a given inquiry. Re-

lating inferences about experimental hypotheses to primary scientific claims is, except in special cases, a distinct step. Yet an additional step is called for to generate data and check the assumptions of the statistical method. Restricting our focus to statistical tests, what I want to consider is how the nonbehavioral construal of tests that I favor supplies answers to two questions: how to specify tests, and how to interpret the results of tests. In so doing, the construal simultaneously answers the main criticisms of NP tests.

Because I think it is important to tie any proposed philosophy of experiment to the actual statistical procedures used, I am deliberately sticking to the usual kinds of test reports—either in terms of the statistical significance of a result, or "accept H" and "reject H"—although it might be felt that some other terms would be more apt. Rather than knock down the edifice of the familiar NP methods, I recommend effecting the nonbehavioral interpretation by setting out rules to be attached to the tests as they presently exist. They might be called "metastatistical" rules. To illustrate, it suffices to consider our by now familiar one-sided test with two hypotheses H and J:

H: μ equals μ_0
J: μ exceeds μ_0.

Parameter μ is the mean value of some quantity, and the experimental statistic for learning about μ is the sample mean, $\overline{X}$. As many of our examples showed, it is often of interest to learn whether observed differences, say in the positive direction, are due to actual discrepancies from some hypothesized value or are typical of chance deviations.

The difference statistic D is the positive difference between the observed mean and the mean hypothesized in H. That is,

$$D = \overline{X} - \mu_0.$$

The NP test, call it T^+, instructs H to be rejected whenever the value of variable $\overline{X}$ differs from H by more than some amount—that is, whenever it exceeds some cutoff point $\overline{X}^*$. One can work with $\overline{X}$ or with D to specify the cutoff point beyond which our test rejects H and accepts J. The cutoff is specified so that the probability of a type I error (rejecting H, given that H is true) is no more than α. Let us suppose that the test T^+ is a "best" NP test with small size α, say, for convenience,[20] that α is .03. Then

20. It is convenient because it corresponds to approximately a 2-standard-deviation cutoff point. If one were looking for discrepancies in both directions, that is, if this were a 2-sided rather than a 1-sided test, then the 2-standard-deviation cutoff would give, approximately, a test with size 0.05. See note 2, chapter 9.

Test T^+: Reject H at level .03 iff $\overline{X}$ exceeds μ_0 by 2 s.d.(x_n)

where s.d.(x_n) is the standard deviation of $\overline{X}$. For simplicity, let the sample size n be large enough to assume approximate normality of $\overline{X}$ according to the central limit theorem (say, n is greater than 30). Let us review the error probabilities of test T^+.

Error Probabilities of T^+

a. The type I error is rejecting H when H is true. The probability of this occurring, α, is no more than .02 because that is the preset size of the test. This holds because $\overline{X}$ exceeds its mean by 2 standard deviations less than 3 percent of the time.

b. The type II error is accepting (or failing to reject) H when H is false (and J is true). Hypothesis J is a "composite" alternative, since it contains all the μ values in excess of μ_0. The probability of the type II error varies depending on which value in J is the true one.

The probability of a type II error is usually written as β, but because it will depend for its value on which specific value in J is true, it is clearer to use $\beta(\mu')$ to refer to the probability of a type II error when μ' is true. One should read $\beta(\mu')$ as follows:

> $\beta(\mu')$: the probability that test T^+ fails to reject H (and accept J) when alternative μ' is true.

The assertion that the mean equals μ' may be written as hypothesis J':

> J': μ equals μ'.

J' is a particular "point" hypothesis within the composite alternative J. That is, $\beta(\mu')$ is the probability of committing a type II error when J' is true. So, $\beta(\mu')$ can also be written

> $\beta(\mu')$: P(test T^+ fails to reject $H \mid J'$ is true).

Notice that "failing to reject H" in test T^+ is equivalent to obtaining an $\overline{X}$ that is not so far from μ_0 as to reach the (2-standard-deviation) cutoff point $\overline{X}^*$. So $\beta(\mu')$ is the probability that $\overline{X}$ is less than $\overline{X}^*$, given that J' is true.

$$\beta(\mu') = P(\overline{X} < \overline{X}^* \mid J' \text{ is true}).$$

As is plausible, test T^+ has a decreasing probability of committing a type II error the "more false" H is—the further μ' is "to the right of" μ_0. One may wish to state this in terms of the complement to the probability of a type II error, namely, the *power* of the test to detect a specific simple alternative μ'. That is,

The power of T^+ *against* $J' = P(\overline{X} \geq \overline{X}^* \mid J' \text{ is true})$.

The power of the test to reject H when J' is true is $1 - \beta(\mu')$. As would be expected, the more discrepant μ' is from μ_0, the higher is test T^+'s power to detect this.

A test's error probabilities may be used to construct arguments from error, arguments based on severity. However, it is important to remember that severity is always calculated relative to a specific hypothesis that a given test passes on the basis of a given outcome. You cannot assess the severity of a test without considering the process of a test's passing a particular hypothesis with one or another outcome of a given experiment. So to relate error probabilities to arguments from error we need to consider specific kinds of inferences that test T^+ can license. We can begin with the simple dichotomy of standard NP tests: positive and negative results.

Positive Results: A Rejection of H

A positive result is the observation of a sample mean that exceeds the hypothesized mean μ_0 by a statistically significant amount. In the standard test T^+ with α set at .03, a statistically significant difference is one that exceeds μ_0 by 2 standard-deviation units. Within the NP model, this is taken as a rejection of H. That is, the cutoff point, $\overline{X}^*$, is $\mu_0 + 2$ s.d.(x_n).

What is learned from an observed difference D_{obs} about the existence of a positive discrepancy from μ_0? For what value of μ' does J: μ exceeds μ' pass a severe test with T^+? From the pattern of arguing from error we get what might be called the rule of rejection (RR):

> RRi. A difference as large as D_{obs} is *a good indication* that μ exceeds μ' just to the extent that it is very probable that test T^+ would have resulted in a smaller difference if μ (the true mean) were as small as μ'.

Notice that this is the same as saying that D_{obs} is *a good indication* that μ exceeds μ' to the extent that J passes a severe test with D_{obs}. That is because "not-J" consists of μ values less than or equal to μ'.[21]

From RRi we get a companion rule for what an observed difference does *not* indicate. Let us set it out separately:

> RRii. D_{obs} is *a poor indication* that μ exceeds μ' if it is very probable that test T^+ yields so large a difference even if μ is no greater than μ'.

21. As discussed in chapter 6, to obtain severity for all of those values it is enough that P (a difference smaller than $D_{obs} \mid \mu$ equals μ') is high. See also the rule of acceptance (RA) below.

The if clause is the same as saying that the claim *J*: μ exceeds μ' fails to pass a severe test (i.e., the probability of not getting so large a difference even if *J* is false is low).

In addition to the rule of rejection (RR), I will be setting out a rule of acceptance (RA). These rules have many uses. They justify the standard prespecified small error probabilities, allow custom-tailoring of inferences after the trial, and serve to avoid common criticisms and misinterpretations of NP tests. Focusing first on rule RR, I will consider each of these uses in turn.

Rule RR Justifies Preset Significance Levels

The concern in the case of rejecting the null or test hypothesis *H* is that a rejection of *H* might be erroneous—that is, the concern is with the type I error. By stipulating that *H* be rejected only if the difference is statistically significant at some small level, it is assured that such a rejection—*at minimum*—warrants hypothesis *J*, that μ exceeds μ_0 by some amount or other. RRi makes this plain.

If *H* is rejected, then the hypothesis that passes the test is *J*, the assertion that μ exceeds the null value μ_0. To obtain severity, we have to consider one minus the probability of such a statistically significant result even if *H* is true (1 − the probability of a type I error). This will vary depending on how much the observed result exceeds the minimal boundary for declaring a result significant enough to reject H_0, namely, $\overline{X}^*$. Its lowest value, however, would be for a result that just makes it to the boundary $\overline{X}^*$. In this case, the observed mean, $\overline{X}_{obs}$, equals the cutoff value $\overline{X}^*$ for calling a result "positive." The severity for this "worst case" of rejecting *H* is one minus the probability of a type I error (i.e., $1 - \beta(\mu_0)$). So the assurance given by a test with a low type I error is that it tells me *ahead of time* that whenever T^+ rejects H_0, *J* has passed a severe test (at least to degree $1 - \alpha$).[22]

22. To review the argument with a bit more detail, remember that a test T^+ with low size or significance level α assures that the cutoff $\overline{X}^*$ beyond which point the sample mean is taken to reject *H* and accept *J* is one that occurs with no more than probability α when *H* is true. That is, it ensures that

1. P(test T^+ yields a sample mean that exceeds $\overline{X}^* \mid H$ is true) $\leq \alpha$.

But (1) ensures that whenever such a test passes *J*, the result is that *J* has passed a severe test, the severity being at least $1 - \alpha$. The reason is that (1) is equivalent to (2):

2. P(test T^+ passes $J \mid H$ is true) $\leq \alpha$.

From which we get

3. P(test T^+ does *not* pass $J \mid H$ is true) $> 1 - \alpha$ (or $\geq 1 - \alpha$ for continuous cases).

And so

4. if *J* passes test T^+, then *J* passes a severe test.

Custom-Tailoring

When outcomes deviate from H by even more than the α cutoff, rule RR justifies two kinds of custom-tailored results: (*a*) it warrants passing J at an even higher severity value than $1 - \alpha$, and (*b*) it warrants passing alternatives J': μ is greater than μ', where μ' is larger than μ_0, with severity $1 - \alpha$. With (*a*) we make the same inference—pass J—but with a higher severity than with a minimally positive result. With (*b*) we keep the same level of severity but make a more informative inference—that the mean exceeds some particular value μ' greater than the null value μ_0.

To illustrate (*b*), suppose that null hypothesis H: $\mu = .5$ in our lady tasting tea example is rejected with a result $\overline{X}_{obs}$ is equal to .7. (Mean μ, recall, is the same as the probability of success p.) That is, out of 100 trials, 70 percent are successes. The 2-standard-deviation cutoff was 60 percent successes, so this result indicates even more than that μ exceeds .5. The result is also a good indication that μ exceeds .6. That is because the observed difference exceeds .6 by 2 standard deviations. The probability of so large a difference from .6 if μ were no greater than .6 is small (about .03). Thus the assertion "μ exceeds .6" passes a severe test with result .7 (severity .97).

Avoiding Misinterpretations and Alleged Paradoxes for NP Tests

Where RRi shows that an α-significant difference from H (for small α) indicates *some* positive discrepancy from H_0, RRii makes it clear that it does *not* indicate any and all positive discrepancies. My failing exam score may indicate that I am ignorant of some of the material yet not indicate that I know none of it.

We can make the point by means of the usual error probabilities, even without customizing to the particular result. Consider the power of a test (1 − the type II error) regarding some alternative μ'. We know:

The power of T^+ against J': $\mu = \mu'$ equals $P(\overline{X} \geq \overline{X}^* \mid J' \text{ is true})$, which equals $1 - \beta(\mu')$. The test's power may be seen as a measure of its sensitivity. The higher the test's probability of detecting a discrepancy μ', the more powerful or sensitive it is at doing so. If, however, a test has a good chance of rejecting H even if μ is no greater than some value μ', then such a rejection is a *poor* indication that μ is even greater than μ'. So—although this may seem odd at first—a statistically significant difference is indicative of a larger discrepancy the *less* sensitive or powerful the test is. If the test rings the alarm (i.e., rejects H_0) even for comparatively tiny discrepancies from the null value, then the ring-

ing alarm is poor grounds for supposing that larger discrepancies exist. As obvious as this reasoning becomes using severity considerations, the exact opposite is assumed in a very common criticism of tests.

Before turning to this criticism, let us illustrate the reasoning in both parts of the RR by means of a medical instrument. Imagine an ultrasound probe to detect ovarian cysts. If the image observed is of the sort that very rarely arises when there is no cyst, but is common with cysts, then the image is a good indication a cyst exists. If, however, you learned that an image of the sort observed very frequently occurred with this probe even for cysts no greater than 2 inches, then you would, rightly, deny that it indicated a cyst as large as, say, 6 inches. The probe's result is a good indication of a cyst of some size, say ¼ inch, but a poor indication of a cyst of some other (much greater) size. And so it is with test results.

If a difference as large as the one observed is very common even if μ equals μ', then the difference does *not* warrant taking μ to exceed μ'. That is because the hypothesis that μ exceeds μ' would thereby have passed a test with poor severity. And in comparing outcomes from two different tests, the one that passes the hypothesis with higher severity gives it the better warrant. How do critics of NP tests get this backwards?

A Fallacy regarding Statistically Significant (Positive) Results

Criticisms of NP tests, we have seen, run to type, and one well-known type of criticism is based on cases of statistically significant (or positive) results with highly sensitive tests. The criticism begins from the fact that any observed difference from the null value, no matter how small, would be classified as statistically significant (at any chosen level of significance) provided the sample size is large enough. (While this fact bears a resemblance to what happens with optional stopping, here the sample size is fixed ahead of time.) There is nothing surprising about this if it is remembered that the standard deviation decreases as the sample size increases. (It is inversely proportional to n.) Indeed, my reason for abbreviating the standard deviation of the sample mean as s.d.($\bar{x}_n$) in this chapter was to emphasize this dependence on n. A 2-standard-deviation difference with a sample size of, say, 100 is larger than a 2-standard-deviation difference with a sample size of 10,000.

We can make out the criticism by reference to a Binomial experiment, such as in the lady tasting tea example. The null hypothesis H is that p, the probability of success on each trial, equals .5. Now, in a sample of 100 trials, the standard deviation of $\bar{X}$ is .05, while in 10,000 trials it is only .005. Accordingly, a result of 70 percent successes is a

very significant result (it exceeds .5 by 4 standard deviations) in a sample of 100 trials. In a sample of 10,000 trials, an equally statistically significant result requires only 52 percent successes! An alleged paradox is that a significance test with large enough sample size rejects the null with outcomes that seem very close to, and by a Bayesian analysis are supportive of, the null hypothesis. This might be called the Jeffreys-Good-Lindley paradox, after those Bayesians who first raised it.

I discuss this paradox at length in Mayo 1985a and elsewhere, but here I just want to show how easy it is to get around a common criticism that is based on it. The criticism of NP tests results only by confusing the import of positive results. The fallacious interpretation results from taking a positive result as indicating a discrepancy beyond that licensed by RR. Howson and Urbach give a version of this criticism (along the lines of an argument in Lindley 1972). Their Binomial example is close enough to the one above to use it to make out their criticism (their p is equal to the proportion of flowering bulbs in a population). The criticism is that in a test with sample size 10,000, the null hypothesis $H: p = .5$ is rejected in favor of an alternative J, that p equals .6 even though .52 is much closer to .5 (the hypothesis being rejected) than it is to .6. And yet, the criticism continues, the large-scale test is presumably a better NP test than the smaller test, since it has a higher power (nearly 1) against the alternative that $p = .6$ than the smaller test (.5).[23]

The authors take this as a criticism of NP tests because "The thesis implicit in the [NP] approach, that a hypothesis may be rejected with increasing confidence or reasonableness as the power of the test increases, is not borne out in the example" (Howson and Urbach 1989, 168). Not only is this thesis not implicit in the NP approach, but it is the exact reverse of the appropriate way of evaluating a positive (i.e., statistically significant) result. The thesis that gives rise to the criticism comes down to thinking that if a test indicates the existence of some discrepancy then it is an even better indication of a very large discrepancy!

Looking at RRii makes this clear. Let us compare the import of the two 4-standard-deviation results, one from a test with sample size $n = 100$, the second from a test with sample size $n = 10{,}000$. In the experiment with 10,000 trials, the observation of 52 percent successes is an extremely *poor* indicator that p is as large as .6. For such a result is very probable even if the true value of p is actually less than .6,

23. I am calculating power here with the cutoff for rejection set at .6—the 2-standard-deviation cutoff.

say, if $p = .55$. Indeed, it is practically certain that such a large result would occur for p as small as .55. Were one to take such a result as warranting that p is .6, one would be wrong with probability very near one.

In contrast, the observation of 70 percent successes with $n = 100$ trials is a very good indication that p is as large as .6. The probability of getting so large a proportion of successes is very small (about .03) if μ is less than .6. The severity of a test that passes "p is as large as .6" with 70 percent successes out of 100 trials is high (.97).

Howson and Urbach's criticism, and a great many others with this same pattern, are based on an error to which researchers have very often fallen prey. The error lies in taking an α-significant difference (from H) with a large sample size as more impressive (better evidence of a discrepancy from H) than one with a smaller sample size.[24] That, in fact, it is the reverse is clearly seen with rule RR. The reasoning can be made out informally with an example such as our ultrasound probe. Take an even more homey example. Consider two smoke detectors. The first is not very sensitive, rarely going off unless the house is fully ablaze. The second is very sensitive: merely burning toast nearly always triggers it. That the first (less sensitive) alarm goes off is a *stronger* indication of the presence of a fire than the second alarm's going off. Likewise, an α-significant result with the *less* powerful test is *more* indicative of a discrepancy from H than with the more powerful test.[25] Interpreting the results accordingly, the authors' criticism disappears.

To be fair, the NP test, if regarded as an automatic "accept-reject" rule, only tells you to construct the best test for a small size α and then accept or reject. A naive use of the NP tools might seem to license the problematic inference. Rule RR is not an explicit part of the usual formulation of tests. Nevertheless, that rule, and the fallacious interpretation it guards against, is part of the error statistician's use of these tests.[26]

24. Rosenthal and Gaito (1963) explain the fallacy as the result of interpreting significance levels—*quite illicitly*—as E-R measures of the plausibility of the null hypothesis. In this view, the smaller the significance level, the less plausible is null hypothesis H, and so the more plausible is its rejection. Coupled with the greater weight typically accorded to experiments as the sample size increases, the fallacy emerges.

25. See Good 1980, 1982 for a Bayesian way of accommodating the diminishing significance of a rejection of H as the sample size increases.

26. The probabilities called for by RR would be obtained using the usual probability tables (e.g., for the Normal distribution). A good way to make use of rule RR without calculating exact severity values for each result is to substitute certain

Negative Results: Failures to Reject

Let us turn now to considering negative results, cases where the observed difference is *not* statistically significant at the specified small α level. Here the null hypothesis H ($\mu = \mu_0$) is not rejected. NP theory describes the result as "accept H," but one must be careful about how to interpret this. As we saw in section 6.5, it would not license the inference that μ is exactly μ_0—that μ does not exceed μ_0 at all. However, as we also saw, we may find a positive discrepancy that *can* be well ruled out. The pattern of reasoning again follows the pattern of arguing from error. We can capsulize this by the following rule of acceptance (RA):

> RAi. A difference as small as D_{obs} is a good indication that μ is less than μ' if and only if it is very probable that a larger difference would have resulted from test T^+ if the mean were as large as μ'.

That is, a statistically *insignificant* difference indicates that J: μ is less than μ' just in case J passes a severe test. As with the RR, we get a companion rule:

> RAii. A difference as small as D_{obs} is a *poor* indication that μ is less than μ' if it is very improbable that the test would have resulted in a larger difference even if the mean were as large as μ'.

Notice that when the result is negative, the error of interest is a false negative (a type II error)—that H will be accepted even though some alternative J is true.

Rule RA Directs Specifying Tests with High Power to Detect Alternatives of Interest

Now T^+ "accepts" H whenever $\overline{X}$ is less than[27] the .03 significance level cutoff. Before the test, one does not yet know what value of $\overline{X}$ will be observed. Ensuring ahead of time that test T^+ has a high power $1 - \beta$ against an alternative J': $\mu = \mu'$ ensures that a failure to reject

---

benchmarks for good and poor indications. Still focusing on test T^+, useful benchmarks for interpreting rejections of hypothesis H would be as follows:

1. A T^+ rejection of H is a *good* indication that μ exceeds $\overline{X}_{obs} - 2\text{s.d.}(x_n)$.
2. A T^+ rejection is a *poor* indication that μ exceeds $\overline{X}_{obs} + 1\text{s.d.}(x_n)$.

(1) corresponds to passing the claim that "μ exceeds $\overline{X}_{obs} - 2\text{s.d.}(x_n)$" with severity .97.

(2) corresponds to passing the claim that "μ exceeds $\overline{X}_{obs} + 1\text{s.d.}(x_n)$" with severity .16. For a more general discussion of benchmarks for both the RR and RA see Mayo 1983.

27. In continuous cases or discrete cases with fairly large n, it does not matter if we take it as $<$ or $\leq$.

H—a case where H passes—is a case that indicates that μ does not exceed μ'. That is, by assuring ahead of time that the power to detect μ' is high, the experimental tester is ensuring that accepting H constitutes passing severely the hypothesis H':

H': μ is no greater than μ'.

Tests should be specified according to the smallest discrepancy from μ_0 that is of interest.

Notice that this power calculation is a calculation of severity for the case where the result just misses the critical boundary for statistical significance. By custom-tailoring this calculation to the particular statistically insignificant result obtained, the after-trial analysis may warrant ruling out values of μ even closer to μ_0.

A variant on this after-trial question is to ask, with regard to a particular alternative μ'', whether the obtained negative result $\overline{X}_{obs}$ warrants ruling out a μ value as large as μ''. Severity tells you to calculate the probability that a mean larger than the one observed, $(\overline{X}_{obs})$, would have occurred, given that the true value of μ were equal to μ''. That is, you must calculate, still referring to test T^+,

$$P(\overline{X} > \overline{X}_{obs} \mid \mu \text{ equals } \mu'').$$

If this value is high, then $\overline{X}_{obs}$ indicates that μ is less than μ''. Equivalently, the claim that μ is less than μ'' passes a severe test with the obtained negative result $\overline{X}_{obs}$. For, were μ as large as μ'', the probability is high that a result greater than the one obtained would have occurred.

11.7 A NOTE ON OBJECTIVITY

The task of specifying the analytical tool for an experimental inquiry (e.g., tests) is a task we placed within the experimental model of our hierarchy. That it lies outside the formalism of standard NP tests has often led critics to charge that NP methods do not really get around the subjectivity that plagues the subjective Bayesian account. Deciding upon test statistics, sample sizes, significance levels, and so on, after all involves judgments—and these judgments, critics allege, are what the NP will "sweep under the carpet" (to use I. J. Good's phrase):

> You usually have to use subjective judgment in laying down your parametric model. Now the *hidebound* objectivist tends to hide that fact; he will not volunteer the information that he uses judgment at all. (Good 1976, 143)

A favorite line of subjective Bayesians is that by quantifying their subjective beliefs they are actually being more objective than users of non-Bayesian, error probability methods. How do we respond to this charge? First, the judgments the NP test requires are not assignments of degrees of belief to hypotheses. Although subjective Bayesians seem to think that all judgments come down to judgments of prior probabilities, I see no reason to accept this Bayesian dogma. Second, there is a tremendous difference between the kinds of judgments in error statistical methods and subjective probability assignments.

The two main differences are these: First, the choice of statistical test may be justified by specific epistemological goals. As rules RR and RA helped us to see, the choice of NP test with low error probabilities reflects a desire to substantiate certain standard types of arguments from error. With increasing experience, experimenters learn which types of tests are likely to provide informative results. There is leeway in the specification, but it is of a rather restricted variety. Often, different studies will deliberately vary test specifications. Indeed, exploiting different ways of analyzing results is often the basis for learning the most. Second, and most important, the latitude that exists in the choice of test does not prevent the determination of what a given result does and does not say. The error probabilistic properties of a test procedure—*however that test was chosen*—allows for an objective interpretation of the results. Let us elaborate on these two points, making reference to the results we have already seen.

Severity and the Epistemological Grounds for Test Specifications

When tests are used in scientific inquiry, the basis for specifying tests reflects the aims of learning from experiment. A low probability of a type I error, for example, is of interest not because of a concern about being wrong some small proportion of times in a long-run series of applications. It is of interest because of what one wants to learn. If you can split off a portion of what you wish to learn so that one of the canonical experimental models can be used, then specifying the test's error properties grows directly out of what one wants to know—what kinds and extents of errors are of interest, what kinds of checks are likely to be available, and so on.

What I am arguing, then, is that the grounds for specifying the error probabilities of tests stem from the experimental argument one wants to be able to sustain. By fixing the type I error at some low value α the experimental tester ensures that any rejection of H, any passing of J, is a good indication that J is the case. It should not be forgotten, of course, that this depends on a suitable choice of distance measure

at step 2 in test construction. In the canonical tests, such as the one just described, the choice of distance measure is already accomplished for us.

But as Neyman and Pearson saw, this leaves too much latitude in the choice of a test. One must also consider the type II error—failing to reject H when H is false. The problem in cases where H is not rejected is that the test may have had little power (probability) of rejecting H even if a discrepancy from H exists. So severity considerations tell us that a failure to reject H cannot be taken as a good indication that H is precisely true, that no discrepancy from H exists. It is, however, possible to find some value of a discrepancy from H that the result "accept H" does warrant ruling out.

What I am proposing, I believe, is a way of drawing out the implications of Pearson's hints and suggestions. Before the trial, we are interested in how to ensure that the experiment is capable of telling us what we want to know, and we set these "worst case" values for the probabilities of type I and type II errors accordingly. After the trial, with the data in hand, Pearson says we should base our conclusions on the actual "tail area" found, which is tantamount to saying, "look at the severity values."

Telling the Truth with Error-Statistics

Of course there is no guarantee that an appropriate test will actually be run. Indeed, the existence of poorly specified and wrongly interpreted NP tests is at the heart of criticisms of that approach. We noted the problem of positive results with too-sensitive tests. An even more common problem arises when negative results arise from too-*insensitive* tests. As A. W. F. Edwards puts it:

> Repeated non-rejection of the null hypothesis is too easily interpreted as indicating its acceptance, so that on the basis of no prior information coupled with little observational data, the null hypothesis is accepted. . . . Far from being an exercise in scientific objectivity, such a procedure is open to the gravest misgivings. (Edwards 1971, 18)

Although such interpretations of negative results occur, it does not follow that they are licensed by the logic of error statistics. They are not. And because researchers must provide us with enough information to assess the error probabilities of their tests, we are able to check if what they want to accept is really warranted by the evidence.

To illustrate, we can pick up on the study on birth-control pills introduced in chapter 5 and scrutinized in section 6.5. The result, recall, was 9 cases of a blood-clotting disorder among women treated

with the birth-control pill compared with 8 out of 5,000 in the control group. Suppose that the researchers reach the following interpretation of their result: "These results indicate that no more than 1 additional case of clotting disorders among 10,000 women on the pill would be expected." That is, using our abbreviation for the risk increase in the population, the researchers infer claim *C*:

C: the evidence indicates H_c: $\Delta < .0001$.

The rule of acceptance (RA) is the basis for denying that this is a warranted interpretation of the results.

The observed difference .0002 was not statistically significant; it reaches a significance level of .4. We can see right away that an observed difference of .0002 or one even more insignificant would occur 50 percent of the time even if the actual increased rate of the disorder was 2 in 10,000.[28] Hence RA tells us that the negative result from this study cannot be taken as ruling out increases as small as 2 in 10,000. The result is just the sort of thing that would occur half the time in studies of substances that cause 2 additional cases of the disorder per 10,000 women. Such an insignificant difference would therefore be even *more* probable if the pill caused only 1 additional case of the disorder in 10,000 women. Hence the result of this study is a poor indication of hypothesis H_c: $\Delta < .0001$. The inference in *C* is not warranted. Such an insignificant result would occur more than half the time even if H_c is false. Equivalently, the assertion H_c passes a test with severity of less than .5, on the basis of this result.

Utilizing a test's error probabilities in this manner, customizing even further to take account of the particular result, enables distinguishing warranted from unwarranted interpretations of the results, and it enables doing so objectively. The objectivity of the assessment is afforded by the objectivity of the error probability properties of the test. Even without calculating precise severity values, we can distinguish (reasonably) warranted and (flagrantly) unwarranted interpretations of results. Plenty of shortcut calculations are available for making this discrimination (see note 26), and more can be developed.[29]

28. This can be seen without any calculations. Label the supposition here as alternative hypothesis *J'*: the increased risk is .0002. Now the observed outcome does not differ at all from what is hypothesized by *J'*. But even if *J'* is true, 50 percent of the time sample differences would be less than .0002, and 50 percent of the time they would be greater. (That is, half of the area under the normal curve would be "to the left of" *J'*, and half "to the right.") See also the discussion of this example in section 6.5. A longer discussion occurs in Mayo 1985b.

29. Consider interpreting negative results, that is, acceptances of *H*, in test T^+. Rule RA directs us to find a value of μ, call it μ^+, such that the result indicates that $\mu < \mu^+$. Equivalently, we are to find the value μ^+ such that the claim "$\mu < \mu^+$"

The latitude in specifying tests is no different from that in the use of other kinds of reliable instruments in science. Understanding the properties of the instruments allows scrutinizing what a given reading does and does not say. The same holds for tests. It does not matter that test specifications might reflect the beliefs, biases, or hopes of the researcher. Perhaps the reason for selecting an insensitive test is your personal desire to find no increased risk, or perhaps it is due to economic or ethical factors. Those factors are entirely irrelevant to scrutinizing what the data do and do not say. They pose no obstacle to my scrutinizing any claims you might make based on the tests, nor to my criticizing your choice of test as inappropriate for given learning goals. There is no sort of comparable basis for criticizing your subjective degrees of belief.

Inferences without Numbers

There is one final objection that may be raised by Bayesians and others wedded to E-R accounts of inference. The present account of testing licenses claims about hypotheses that are and are not indicated by tests without assigning quantitative measures of support or probability to those hypotheses. But without such assignments of support or probability to hypotheses, the E-R theorist, I expect, will deny that the present account constitutes a genuine account of inductive or statistical inference. Yet this is just to assume that an E-R account is what is needed, and that is what those who embrace testing accounts of inference wish to deny. The Bayesian critic may persist that if I do not secretly really mean to assign some number to the inferences licensed by my tests, then what do I mean by evidence indicating hypotheses? My answer is the one I have been giving throughout this book. That data indicate hypothesis H means that the data indicate or signal that H is

---

passes a severe test. Say we take .97 as a benchmark for severity. Then μ^+ would equal $\overline{X}_{obs} + 2\text{s.d.}(x_n)$. (See also section 6.5.)

Mathematically, the calculation of μ^+ (for the case of test T^+) is equivalent to formulating the *upper confidence bound* of a (one-sided) interval estimate at the corresponding level of confidence. However, unlike the report that "μ is somewhere between μ_0 and μ^+," RA instructs a distinct severity assessment for each value in the interval. More generally, RA directs us to understand what a specific negative statistical result indicates (more or less well) by calculating all or several of the upper bounds for different degrees of severity. This would yield what might be called *severity curves*. It most closely corresponds to forming a series of upper confidence intervals, one for each confidence level. I have recently come across an article by Poole (1987) using what are essentially severity curves in medical statistics. Similar curves are employed by Kempthorne and Folks (1971), but with a different interpretation. Clearly, more work is called for in studying statistical practice and in generalizing these ideas.

correct—much as I might say that a scale reading indicates my weight. Generally, several checks of a given indication of H (e.g., checks of the experimental assumptions) are required before reaching the inference that the data indicate the correctness of H. What does it mean to infer that H is indicated by the data? It means that the data provide good grounds for the correctness of H—good grounds that H correctly describes some aspect of an experimental process. What aspect, of course, depends on the particular hypothesis H in question. One can, if one likes, construe the correctness of H in terms of H being reliable, provided one is careful in the latter's interpretation. Learning that hypothesis H is reliable, I proposed (chapter 4), means learning that what H says about certain experimental results will often be close to the results that would actually be produced—that H will or would often succeed in specified experimental applications. What further substantive claims are warranted will depend on the case at hand.

What is learned receives a formal construal in terms of experimental distributions—assertions about what outcomes would be expected, and how often, if certain experiments were to be carried out. Informally and substantively, this corresponds to learning that data do or do not license ruling out certain errors and mistakes.

To those who insist that every uncertain inference must have a quantity attached, our position is that this insistence is seriously at odds with the kinds of inferences made every day, in science and in our daily lives. There is no assignment of probabilities to the claims themselves when we say things such as the evidence is a good (or a poor) indication that light passing near the sun is deflected, that treatment X prolongs the lives of AIDS patients, that certain dinosaurs were warm blooded, that my four-year-old can read, that metabolism slows down when one ingests fewer calories, or any of the other claims that we daily substantiate from evidence.

Concluding Remarks

To summarize, the key difference between standard NP methods and those based on the likelihood principle is that the former have an interest in and an ability to control error probabilities, whereas the latter do not. Criticisms of NP tests that are not merely misinterpretations arise from supposing that long-run error probabilities are all that matter in NP tests, and that the reason error probabilities matter in NP tests is their interest in ensuring a low probability of erroneous "acts" in the long run. A Pearsonian error statistician denies both of these suppositions. For a Pearsonian, the ability to control error probabilities matters (in a scientific context) because of the desire to correctly learn

about underlying causes, distinguish genuine from spurious effects, and so on, to all that may be learned by arguing from error.

On the Pearsonian view of tests, the greater "seriousness" the behavioristic model attaches to the type I error goes over into the concern to be assured that a rejection of H is a good indication of the existence of a real departure from H, for example, a real effect. The particular balance chosen between the two types of errors is not an arbitrary matter reflecting pragmatic, decision-theoretic values, as Fisher had feared. In learning contexts, their specification is guided by the aims of inquiry, by what one wants to learn. After the results are in, utilizing these error probabilities is the key to scrutinizing objectively inferences based on test results.

In any substantive inquiry, NP methods would need to be used for a series of tests aimed at rejecting different types of alternatives and errors. Rejecting a "chance" hypothesis H, with its indication that some systematic factor is operating, is likely to be only a first step. Ruling out other substantive factors may be accomplished with subsequent statistical tests linking different experimental and data models. 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 understanding how standard error statistical methods afford this type of *piecemeal* approach that one can capture the manner in which these tools are used in day-to-day experimental inquiries.

However, Pearson's advocacy of a piecemeal, inferential use of NP tests requires him to reject the basic tenets of the behavioral decision philosophy that has come to be associated with NP methods. There is no inconsistency in his rejection. While the interpretation of test results differs from the behavioral-decision one, still retained is what is central to error statistical theory: the focus on a procedure's error probabilities. The control of error probabilities has fundamental uses in learning contexts. The link between controlling error probabilities and experimental learning comes by way of the link between error probabilities and severity. The ability to provide methods whose actual error probabilities will be close to those specified by a formal statistical model, I believe, is the key to achieving experimental knowledge.