4 An Error-Statistical Philosophy of Evidence

Deborah G. Mayo

ABSTRACT

Despite the widespread use of error-statistical methods in science, these methods have been the subject of enormous criticism, giving rise to the popular statistical "reform" movement and bolstering subjective Bayesian philosophy of science. Given the new emphasis of philosophers of science on scientific practice, it is surprising to find they are rarely called upon to shed light on the large literature now arising from debates about these reforms—debates that are so often philosophical. I have long proposed reinterpreting standard statistical tests as tools for obtaining experimental knowledge. In my account of testing, data x are evidence for a hypothesis H to the extent that H passes a severe test with x. The familiar statistical hypotheses as I see them serve to ask questions about the presence of key errors: mistaking real effects for chance, or mistakes about parameter values, causes, and experimental assumptions. An experimental result is a good indication that an error is absent if there is a very high probability that the error would have been detected if it existed, and yet it was not detected. Test results provide a good (poor) indication of a hypothesis H to the extent that H passes a test with high (low) severity. Tests with low error probabilities are justified by the corresponding reasoning for hypotheses that pass severe tests.

Is it possible to have a general account of scientific evidence and inference that shows how we learn from experiment despite uncertainty and error? One way that philosophers have attempted to affirmatively answer this question is to erect accounts of scientific inference or testing where appealing to probabilistic or statistical ideas would accommodate the uncertainties and error. Leading attempts take the form of rules or logics relating evidence (or evidence statements) and hypotheses by measures of confirmation, support, or probability. We can call such accounts *logics of evidential relation-*

ship (or *E-R logics*). In this view, philosophers could help settle scientific disputes about evidence by developing and applying logics of evidential relationship. I will begin my discussion by reflecting on these logics of evidence and how they differ from what I think an adequate account of evidence requires.

LOGICS OF EVIDENTIAL RELATIONSHIP VS. ERROR STATISTICS

The leading example of a logic of evidential relationship is based on one or another *Bayesian* account or model. In the most well-known of Bayesian accounts, our inference about a hypothesis H from evidence e is measured by the probability of H given e using Bayes' theorem from probability theory. Beginning with an accepted statement of evidence e, scientists are to assign probabilities to an exhaustive set of hypotheses. Evidence e *confirms* or supports hypothesis H to the extent that the probability of H given the evidence exceeds the initial assignment of probability in H. But how can we make these initial assignments of probability (i.e., the prior probabilities)? In early attempts at building an "inductive logic" (by Carnap and other logical empiricists), assignments based on intuitive, logical principles were sought, but paradoxes and incompatible measures of evidential relationship resulted. Moreover, the inductive logicians were never able to satisfactorily answer the question of how purely logical measures of probability are to be relevant for predictions about actual experiments. Contemporary subjective Bayesian philosophers instead interpret probabilities as an agent's subjective degrees of belief in the various hypotheses. The resulting *subjective Bayesian way* furnishes the kind of logic of evidential relationship that many philosophers have sought.

Although the subjective Bayesian approach is highly popular among philosophers, its dependence upon subjective degrees of belief, many feel, makes it ill-suited for building an objective methodology for science. In science, it seems, we want to know what the data are saying, quite apart from the opinions we start out with. In trading logical probabilities for measures of belief, the problem of relevance to real world predictions remains. By and large, subjective Bayesians admit as much. Leonard (L. J.) Savage, a founder of modern "personalistic" Bayesianism, makes it very clear throughout his work that the theory of personal probability "is a code of consistency for the person applying it, not a system of predictions about the world around him" (Savage, 1972, 59). But a code for consistently adjusting subjective beliefs is unable to serve as the basis for a philosophy of evidence that can help us to understand, let alone adjudicate objectively, disputes about evidence in science.

Under E-R logics, one might wish to include the less familiar *likelihood* approach (or approaches). The likelihoodist uses the likelihood ratio as the comparative evidential-relation measure.[1] I shall have more to say about this approach as we proceed.

A second statistical methodology on which one might erect a philosophy of evidence is alternatively referred to as classical, orthodox, frequentist, or Neyman-Pearson (NP) statistics. The methods and models of NP statistics (e.g., statistical significance tests, confidence interval estimation methods) were deliberately designed so that their validity does not depend upon prior probabilities to hypotheses—probabilities that are eschewed altogether except where they may be based upon objective frequencies. Probability arises not to assign degrees of belief or confirmation to hypotheses but rather to characterize the experimental testing process itself: to express how *frequently* it is capable of discriminating between alternative hypotheses and how *reliably* it facilitates the detection of error. These probabilistic properties of experimental procedures are called *error frequencies* or *error probabilities* (e.g., significance levels, confidence levels). Because of the centrality of the role of error probabilities, NP statistics is an example of a broader category that I call *error probability statistics* or just *error statistics*.

Measures of Fit vs. Fit + Error Probabilities

A key feature distinguishing error-statistical accounts from evidential-relation logics (whether Bayesian, likelihoodist, hypothetico-deductive, or other) is that the former requires not just a measure of how well data "fit" hypotheses but also a calculation of the error probabilities associated with the fit. If you report your E-R measure of fit, the error statistician still needs to know how often such a good fit (or one even better) would arise even if the hypothesis in question were false—an error probability. For a NP error statistician, two pieces of evidence that fit a given hypothesis equally well may have very different evidential imports because of a difference in the er-

1. The so-called law of likelihood asserts that evidence x supports hypothesis H more than hypothesis J if the likelihood of H exceeds that of J. Leaders in developing a likelihood account (Hacking, Barnard, Birnbaum) have abandoned or greatly qualified their acceptance of this approach. Some likelihoodists also include prior probabilities (e.g., Edwards). Current-day likelihoodist accounts are discussed in Berger and Wolpert (1988) and Royall (1997, 2004).

ror probabilities of the procedures from which each data set was generated. For Bayesians, by contrast, because such a distinction is based on considering outcomes other than the one actually observed, it leads to "incoherence." That is because Bayesians, and also likelihoodists, accept what is called the *likelihood principle*.[2] As Edwards, Lindman, and Savage (1963, 238) put it, "Those who do not accept the likelihood principle believe that the probabilities of sequences that might have occurred, but did not, somehow affect the import of the sequence that did occur." We error statisticians do indeed: error probability considerations always refer to outcomes other than the ones observed, and such considerations are at the foundation for scrutinizing evidence. For a full discussion, see Mayo and Kruse (2001).

This distinction about the role of error probabilities is closely linked to a difference in attitude about the importance, in reasoning from data, of considering how the data were generated. E. S. Pearson puts it plainly,

> *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 [error probabilities] 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.* (Pearson, 1966, 277–278; emphasis added)

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

> The Bayesian theory of support is a theory of how the *acceptance as true of some evidential statement* affects your belief in some hypothesis. How you came to accept the truth of the evidence, and whether you are correct in accepting it as true, are matters that, from the point of view of the theory, are simply irrelevant, (Howson and Urbach, 1993, 419; emphasis added)

2. Following Edwards, Lindman, and Savage (1963), the likelihood principle may be stated by considering two experiments with the same set of hypotheses H_1 up to H_n. If D is an outcome from the first experiment and D′ from the second, then D and D′ are evidentially equivalent when $P(\mathrm{D}'; H_i) = P(\mathrm{D}; H_i)$ for each i.

The presumption that philosophical accounts of hypothesis appraisal begin their work with some statement of evidence e as given or as accepted is a key feature of logics of evidential relationship. This holdover from logical empiricist philosophies—where statements of evidence were regarded as relatively unproblematic "givens"—is a central reason that philosophers have failed to provide accounts of evidence that are relevant for the understanding and elucidating of actual scientific practice (Mayo, 2003a).

What We Really Need in a Philosophy of Evidence

In practice, where data are inexact, noisy, and incomplete, we need an account to grapple with the problems in determining if we even have evidence for a hypothesis to begin with. Thus, an adequate account must not begin with given statements of evidence but should include questions about how to generate, model, use, or discard data, and criteria for evaluating data. Even where there are no problems with the data, we need an account that makes sense of the fact that there is often disagreement and controversy as to whether data provide evidence for (or against) a claim or hypothesis of interest. An account that spews out an assessment of evidential support whenever we input statements of evidence and hypotheses does not do that. Settling or at least making progress with disputes over evidence requires appealing to considerations that are disregarded in Bayesian and other E-R accounts: considerations of reliability understood as error probabilities. Two pieces of evidence that would equally well warrant a given hypothesis according to logical measures of evidential relationship may in practice be regarded as differing greatly in their evidential value because of differences in how reliably each was produced. More specifically, as I see it, scientists seek to scrutinize whether the overall experiment from which the data arose was a reliable probe of the ways we could be mistaken in taking $\boldsymbol{x}$ as evidence for (or against) H. Scientists seem willing to forgo grand and unified schemes for relating their beliefs in exchange for a hodgepodge of methods that offer some protection against being misled by their beliefs.

This does not mean we have to give up saying anything systematic and general—as many philosophers seem to think. The hodgepodge of methods give way to rather neat statistical strategies of generating and interpreting data, and the place to look for organizing and structuring such activities, I maintain, is the host of methods and models from standard error statistics, error analysis, experimental design, and cognate methods.

Despite the increasingly widespread use of error-statistical methods in science, however, they have been the subjects of continuing philosophical

controversy. Given the new emphasis philosophers of science have placed on taking their cues from actual scientific practice, it is disappointing to find them either ignoring or taking sides against error-statistical methods rather than trying to resolve these controversies, or at least helping to clarify the reasoning underlying probabilistic procedures that are so widely used to assess evidence in science. Increased availability of sophisticated statistical software has made the disputes in philosophy of statistics more pressing than ever; they are at the heart of several scientific and policy controversies—and philosophers can and should play a role in clarifying if not resolving them.

CRITICISMS OF NP STATISTICS AND THEIR SOURCES

While the confusion and controversy about NP tests have generated an immense literature, the traditional criticisms run to type. (One may identify three generations of philosophy of statistics disputes. See Mayo and Spanos, 2004.) They may be seen to originate from two premises, both false. The first deals with assumptions about what NP methods can provide, the second with assumptions about "what we really want" from a theory of statistical evidence and inference. The first set of assumptions, about what NP methods can give us, is based on what may be called the *behavioral-decision* interpretation or model of NP tests. This model regards NP tests as providing mechanical rules or "recipes" for deciding how to "act" so as to ensure that one will not behave "erroneously" too often in the long run of experience. The second assumption, about "what we really want," reflects the image of inductive inference as providing an E-R logic. It is assumed, in other words, that inductive inference must take the form of providing a final quantitative measure of the absolute or relative support or probability of hypotheses given data. But the only probabilistic quantities the NP methods provide are error probabilities of procedures. Unsurprisingly, as critics show, if error probabilities (e.g., significance levels) are interpreted as probabilities of hypotheses, misleading and contradictory conclusions are easy to generate. Such demonstrations are not really criticisms but rather are flagrant misinterpretations of NP methods.

Not all criticisms can be dismissed so readily. Other criticisms proceed by articulating criteria for testing or for appraising evidence thought to be intuitively plausible and then showing that tests that are good according to NP error criteria may not be good according to these newly defined criteria

or principles. Such criticisms, although they can be subtle and persuasive, are guilty of begging the question against NP principles[3] (see Mayo, 1985a).

This type of criticism takes the following form: NP procedures cannot be used to interpret data as evidence because it leads to different results even when the evidence is the same.[4] This sounds like a devastating criticism until one looks more closely at the concept of "same evidence" that is being assumed. What the charge really amounts to is this: NP methods can lead to distinct interpretations of evidence that would not be distinguished on the basis of measures of evidential relationship that regard error probabilities as irrelevant. But NP procedures do regard differences in the error probabilities from which data arose as relevant for interpreting the evidential import of data. Hence, what at first blush appears to be a damning criticism of NP theory turns out to be guilty of begging the question against a fundamental NP principle.

But several critics seem to think that one cannot live within the strictures of NP principles, and several have argued that researchers invariably use NP tests in an inconsistent fashion that reflects "a critical defect in current theories of statistics" (Royall, 1997, xi). This common criticism has even been given a colorful Freudian twist as discussed by Gerd Gigerenzer (1993). According to Gigerenzer, researchers who use NP tests actually use a hybrid logic mixing ideas from NP tests, Fisherian tests, and Bayesian analysis—leading to a kind of Freudian "bad faith." In Gigerenzer's "Freudian metaphor," "the NP logic of hypothesis testing functions as the Superego of the hybrid logic" (Gigerenzer, 1993, 324). It insists upon prespecifying significance levels, power, and precise alternative hypotheses and then accepting or rejecting the null hypotheses according to whether outcomes fall in rejection regions of suitable NP tests. The Ego—the fellow who gets things done—is more of a Fisherian tester, we are told. The Fisherian Ego does not use prespecified significance levels but reports attained *P*-values after the experiment. He likewise ignores power and the alternative hypotheses altogether—despite their being an integral part of the NP methodology. The Ego violates the NP canon further by allowing himself epistemic statements

3. This is so, at any rate, for all criticisms of which I am aware. There are also criticisms (generally raised by Bayesians) that argue that NP methods fail to promote their own goal of low error probabilities. I discuss this in my rejoinder.

4. Richard Royall states it plainly: "Neyman-Pearson theory leads to different results in two situations where the evidence is the same, and in applications where the purpose of the statistical analysis is to represent and interpret the data as evidence, this is unacceptable" (Royall, 1997, 47).

that go beyond those for which the NP test (regarded as a mechanical decision rule) strictly permits; and in so doing, the Ego "is left with feelings of guilt and shame having violated the rules" (325). Finally, we get to the Bayesian Id. "Censored by both the frequentist Superego and the pragmatic Ego are statements about probabilities of hypotheses given data. These form the Bayesian Id of the hybrid logic," says Gigerenzer (325). The metaphor, he thinks, explains "the anxiety and guilt, the compulsive and ritualistic behavior, and the dogmatic blindness" associated with the use of NP tests in science. (Perhaps some might replace the Bayesian Id with the likelihoodist Id.)

Doubtless Gigerenzer is correct in claiming that statistics texts erroneously omit these philosophical and historical differences between NP tests, Fisherian tests, and Bayesian methods, and doubtless many have been taught statistics badly, but we need not turn this into an inevitability. What we need (and what I have elsewhere tried to provide) is an interpretation of statistical tests that shows how they can be seen to produce a genuine account of evidence without misinterpreting error probabilities and without being used as mechanical "cookbook" methods for outputting "acts" associated with "accept *H*" or "reject *H*."

BEYOND THE BEHAVIORAL-DECISION MODEL OF NP TESTS

The place to begin is with the first premise upon which these assaults on NP tests are based—the assumed behavioral decision model of tests. Granted, the proof by Neyman and Pearson of the existence of "best" tests, and Wald's later extension of NP theory into a more generalized decision model, encouraged the view that tests (particularly best tests) provide the scientist with a kind of *automatic rule* for deciding to accept or reject hypotheses. Nevertheless, the whole concept of "inductive behavior," a term coined by Neyman, was just Neyman's way of distinguishing these tests from the concept of "inductive inference" as understood by Bayesians as well as by Fisherians. Even that arch opponent of Neyman, subjectivist Bruno de Finetti, remarked that the expression "inductive behavior . . . that was for Neyman simply a slogan underlining and explaining the difference between his, the Bayesian and the Fisherian formulations" became, with Abraham Wald's work, "something much more substantial" (de Finetti, 1972, 176). De Finetti called this "the involuntarily destructive aspect of Wald's work" (ibid.). Pearson made it clear that he, at least, had always intended an "evidential" and

not a behavioral interpretation of tests. To the statistician Allan Birnbaum, for instance, Pearson wrote 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. (Birnbaum, 1977, 33; emphasis added)

Pearson explicitly distanced himself from the inductive behavior concept, telling R. A. Fisher that inductive behavior "is Professor Neyman's field rather than mine" (Pearson, 1955, 207), and declaring that "no responsible statistician . . . would follow an automatic probability rule" (Pearson 1947, 192).

But even Neyman used tests to appraise evidence and reach conclusions; he never intended the result of a statistical test to be taken as a decision about a substantive scientific hypothesis. Addressing this point, Neyman declared, "I do not hesitate to use the words 'decision' or 'conclusion' every time they come handy" (Neyman, 1976, 750). He illustrates by discussing a test on a null hypotheses of no difference in means: "As a result of the tests we applied, we decided to act on the assumption (or concluded) that the two groups are not random samples from the same population" (ibid.). What needs to be understood is that Neyman regards such a decision as one of how to appraise the data for the purpose of subsequent inferences in an ongoing inquiry.[5] See also Neyman (1955). The tendency of philosophers to seek final assessments of hypotheses and theories leads them to overlook the fact that in practice evidential appraisal is part of an ongoing inquiry, and that this calls for tools that help guide us in what to try next and how to communicate data in forms that will allow others to critique and extend results. Viewed as part of a piecemeal tool for ongoing inquiry, a very different picture of the value of NP tools emerges. (See Mayo, 1990, 1996.)

Nevertheless, my sympathies are much closer to Pearson's, who would have preferred that the tests be articulated in the evidential manner in which he claims they were first intended—namely as "a means of learning" (Pearson, 1955, 206). Still, the behavioral-decision concepts, it seems to me, can be regarded as simply a way to characterize the key formal features of

5. Here Neyman concludes the data do not warrant the assumption that the two groups are random samples from the same population—a conclusion that then serves as input for the primary inquiry at hand.

NP tools, and understanding these features enables tests to perform the nonbehavioral tasks to which tests may be put to learn from evidence in science. The task remains to explicate the nonbehavioral or evidential construal of these methods. While the reinterpretation I propose differs sufficiently from the standard NP model to warrant some new name, it retains the centerpiece of these methods: the fundamental use of error probabilities—hence the term *error statistics.*

A FRAMEWORK OF INQUIRY

To get at the use of these methods in learning from evidence in science, I propose that experimental inference be understood within a framework of inquiry. You cannot just throw some "evidence" at the error statistician and expect an informative answer to the question of what hypothesis it warrants. A framework of inquiry incorporates methods of experimental design, data generation, modeling, and testing. For each experimental inquiry we can delineate three types of models: *models of primary scientific hypotheses, models of data,* and *models of experiment.* Figure 4.1 gives a schematic representation.

A substantive scientific inquiry is to be broken down into one or more local hypotheses that make up the *primary questions* or *primary problems* of distinct inquiries. Typically, primary problems take the form of estimating quantities of a model or theory, or of testing hypothesized values of these quantities. The *experimental models* serve as the key linkage models connecting the primary model to the data—links that require not the raw data itself but appropriately *modeled data.*

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 primary scientific claims is a distinct step except in special cases. Yet a third step is called for to link raw data to data models—the real material of experimental inference. The indirect and piecemeal nature of our use of sta-

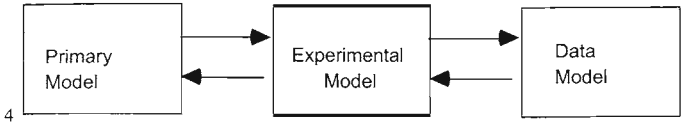

FIGURE 4.1 Models of Experimental Inquiry.

tistical methods, far from introducing an undesirable complexity into our approach, is what enables it to serve as an account of inference that is relevant to scientific practice.[6]

SEVERE TESTS AND ARGUING FROM ERROR

Despite the complexity, there is an overall logic of experimental inference that emerges: Data x *indicate* the correctness of hypothesis H, to the extent that H passes a *severe test* with x. Hypothesis H passes a severe test with x if (a) x fits H (for a suitable notion of fit)[7] and (b) the test procedure had a high probability of producing a result that accords less well with H than x does if H were false or incorrect. To infer that H is indicated by the data does not mean a high degree of probability is assigned to H—no such probabilities are wanted or needed in the error-statistical account. That data indicate that hypothesis H means that the data indicate or provide reliable evidence that hypothesis H is correct—good grounds that H correctly describes some aspect of an experimental process. What aspect, of course, depends on the particular hypothesis H in question. Generally, several checks of a given indication of H (e.g., checks of the experimental assumptions) are required. 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 H is reliable would mean learning that what H says about certain experimental results will or would often be close to the results that would actually be produced—that H will or would often succeed in specified experimental applications.[8] This is *experimental knowledge.*

The reasoning used in arriving at this knowledge follows an informal pattern of argument that I call *an argument from error* or *learning from error.* The overarching structure of the argument is guided by the following thesis:

6. For much further discussion and development of the error-statistical account, see Mayo (1996).

7. I leave this notion open in order to accommodate different inference accounts. In some accounts, x fits H to the extent that $P(x; H)$ is high. Because $P(x; H)$ is often small, even if H is true, x might be said to fit H if x is more probable under H than under all (or certain specified) alternatives to H.

8. What is learned may be formally construed in terms of experimental distributions—assertions about what outcomes would be expected, and how often, if certain experiments were to be carried out. Informally, one can see this as corresponding to learning that data do or do not license ruling out certain errors and mistakes.

> It is learned that an error is absent [present] when (and only to the extent that) a procedure of inquiry (which may include several tests) with a high probability of [not] detecting the error if and only if it is present [absent], nevertheless does not [does] detect the error.[9]

Not detecting the error means it produces a result (or set of results) that is in accordance with the absence of the error. Such a procedure of inquiry may be called a reliable (or highly severe) error probe. According to the above thesis, we can argue that an error is absent if it fails to be detected by a highly reliable error probe. (A corresponding assertion as to the error's absence may be said to have passed a severe test.)

It is important to stress that my notion of severity always attaches to a particular hypothesis passed or a particular inference reached. How severe is this test? is not a fully specified question until it becomes, how severe would a test procedure be, if it passed such-and-such hypothesis on the basis of such-and-such data? A procedure may be highly severe for arriving at one type of hypotheses and not another.[10] Also, severity is not equal to the formal notion of *power:* the two would give identical measures only if one was testing a (continuous) statistical hypothesis H and obtained an outcome that just misses the cutoff for rejecting H.[11]

THE ROLES OF STATISTICAL MODELS AND METHODS

Experimental inquiry is a matter of building up, correcting, and filling out the models needed for substantiating severe tests in a step-by-step manner.

9. In terms of a hypothesis H, the argument from error may be construed as follows: Evidence in accordance with hypothesis H indicates the correctness of H when (and only to the extent that) the evidence results from a procedure that with high probability would have produced a result more discordant from H, were H incorrect.

10. To illustrate, consider a diagnostic tool with an extremely high chance of detecting a disease. Finding no disease (a clean bill of health) may be seen as passing hypothesis H_1: no disease is present. If H_1 passes with so sensitive a probe, then H_1 passes a severe test. However, the probe may be so sensitive as to have a high probability of declaring the presence of the disease even if no disease exists. Declaring the presence of the disease may be seen as passing hypothesis H_2: the disease is present. If H_2 passes a test with such a highly sensitive probe, then H_2 has not passed a severe test. That is because there is a very low probability of not passing H_2 (not declaring the presence of the disease) even when H_2 is false (and the disease is absent). The severity of the test that hypothesis H_2 passes is very low. This must not be confused with a posterior probability assignment to H_2. See Mayo, 2004, and chapter 4.3 in this volume.

11. The severity with which "accept H" passes would, in this case, be equal to (or nearly equal to) the power of the test.

Error-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. They fulfill three main tasks by providing:

1. canonical models of low-level questions with associated tests and data-modeling techniques;
2. tests and estimation methods that allow control of error probabilities;
3. techniques of data generation and modeling along with tests for checking whether the assumptions of data models are met.

The three tasks just listed relate to the primary models, the models of experiment, and of data, respectively. In each case, I readily admit that the functions served by the statistical tools do not fall out directly from the mathematical framework found in statistical textbooks. There are important gaps that need to be filled in by the methodologist and philosopher of experiment.

Task (i). This is especially so for the first task, providing models of low-level questions. Explicating this task, as I see it, enjoins us to ask how scientists break down a substantive inquiry into piecemeal questions such that they can be reliably probed by statistical tests or analogs to those tests. The questions, I propose, may be seen to refer to standard types of errors. Four such standard or canonical types of errors are:

1. mistaking chance effects or spurious correlations for genuine correlations or regularities;
2. mistakes about a quantity or value of a parameter;
3. mistakes about a causal factor;
4. mistakes about the assumptions of experimental data.

Statistical models are relevant because they model patterns of irregularity that are useful for studying these errors.

Task (ii). The second task centers on what is typically regarded as statistical inference proper; namely, specifying and carrying out statistical tests (and the associated estimation procedures) or informal analogs to these tests. It is important to emphasize, however, that the error-statistical program brings with it reinterpretations of NP tests as well as extensions of their logic into informal arguments from error. The criteria for selecting tests depart from those found in classic behavioristic models of testing. One seeks not the "best" test according to the low error probability criteria alone, but rather sufficiently informative tests.

What directs the choice of a test reflects not merely a concern to control a test's error probabilities but also a desire to ensure that a test is based on a plausible "fit" or *distance measure* (between data and hypotheses). The recognition of these twin concerns allows answering a number of criticisms of NP tests. While I admit this takes us beyond the usual formulations of NP tests, there is plenty of evidence that Pearson, at least, intended such features 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, 1947, 173), Pearson defines three steps in specifying tests:

> Step 1 We must specify the experimental probability set [the sample space, discussed in chapter 1 of this volume)] 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, 1947, 173)

Results make us "more and more inclined" to reject H as they get further away from the results expected under H; that is, as the results become more probable under the assumption that some alternative J is true. 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 (Pearson, 1947, 173).[12]

For example, step 2 might give us the likelihood or the ratio of likelihoods of hypotheses given evidence, and at step 3 we may calculate the probability of getting so high a likelihood ratio, say in favor of J against hypothesis H, when in fact H is true.

12. 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.

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, he explains, did formulations of the NP model begin first by fixing the error value for step 3 and then determining the associated critical bounds for the rejection region.[13]

If the rationale were solely error probabilities in the long run, the need to first deliberate over an appropriate choice of measuring distance at step 2 would indeed drop out, as in the behavioral-decision formulation. 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. From step 3 we can calculate how the test, selected for its error probabilities, is dividing the possible outcomes. But this is different from having first deliberated at step 2 as to which outcomes are "further from" or "closer to" H in some sense, and thereby should incline us more or less to reject H. The resulting test, while having low error probabilities, may fail to ensure that the test has an increasing chance of rejecting H the more the actual situation deviates from the one that H hypothesizes. The resulting test may even be irrelevant to the hypothesis of interest. The reason that many counterintuitive tests appear to be licensed by NP methods, e.g., certain mixed tests,[14] is that tests are couched in the behavioral framework in which the task Pearson intended for step 2 is absent.[15]

Task (iii). The third task involves pretrial planning to generate data likely to justify needed assumptions, and after-trial checking to test if the assumptions are satisfactorily met. Often this is accomplished by deliberately

13. 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" (Pearson, 1947, 173).

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 do Pearson and "error statisticians." An error statistician 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.

15. Exceptions exist, e.g., the exposition of tests in Cox and Hinkley (1974) and in Kempthorne and Folks (1971), in which test statistics are explicitly framed in terms of distance measures.

introducing statistical considerations (e.g., via randomization). We noted earlier how Pearson linked the importance of error probabilities to the fact that the NP account regards the "preliminary planning and subsequent interpretation" as closely linked. 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 one what one wants to learn; for example, one can determine ahead of time how large a sample would need to be in a certain test to have a reasonably high chance (power) of rejecting H when in fact some alternative J is true. Few dispute this (before-trial) function of error probabilities, it seems.

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 interpret correctly 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. It is on this after-trial function that I want to focus; for this is what non-error-statistical approaches (those accepting the likelihood principle) deny.[16]

Although there are several (after-trial) uses of error probabilities, they may all be traced to the fact that error probabilities are properties of the procedure that generated the experimental result. 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 whether given claims are or are not mistaken descriptions of the data-generating procedure. It teaches us how typical (or how rare) given results would be under varying hypotheses about the experimental process. A central use of this information is to determine what it would be like if various hypotheses about the underlying experimental process misdescribed a specific experimental process. (It teaches us what it would be like were it a mistake to suppose a given effect were nonsystematic or due to chance, what it would be like were it a mistake to attribute the effect to a given factor, what it would be like were it a mistake to hold that a given quantity or parameter had a certain value, and what it would be like were it

16. Some (e.g., Hacking, 1965) have suggested that error probabilities, while acceptable for before trial planning, should be replaced with other measures (e.g., likelihoods) after the trial. Pearson (1966) takes up and rejects this proposal, raised by Barnard in 1950, reasoning that "if the planning . . . is based on a study of the power function of a test and then, having obtained our results, we do not follow the first rule but another, based on likelihoods, what is the meaning of the planning?" (Pearson, 1966, 228).

a mistake to suppose experimental assumptions are satisfactorily met.) Statistical tests can then be designed so as to magnify the differences between what it would be like under various hypotheses.

The need for such information develops out of an awareness that we can easily be misled if we look only at how well data fit hypotheses, ignoring sampling rules (e.g., rules for when to stop collecting data or *stopping rules*) and failing to take into account whether hypotheses were deliberately selected for testing because they fit the data (hypothesis specification). Although these features of sampling and specifying hypotheses do not alter measures of fit such as likelihood ratios, they do alter error probabilities. So requiring a report of error probabilities alerts us when an altered appraisal of the evidence is required. That is why fitting, even being the best-fitting hypothesis, is not enough for the error statistician. Pearson and Neyman (1930) 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. (Pearson and Neyman, 1930, 106)[17]

Examples of how likelihood ratios alone fail to control error probabilities are well known. One type of case is where hypothesis H is the common null hypothesis that asserts an observed correlation e results from mere "chance." (H asserts, in effect, that it would be an error to regard e as the result of a genuine rather than a chance correlation.) Suppose that $P(e; H)$ is small. Since we can always find an alternative hypothesis J that perfectly fits the data (so the likelihood of J is maximal), we can always get a high likelihood ratio in favor of J against H. Because one could always do this even when H is true, the probability of finding support for J against H *erroneously* is 1. (One always denies the error asserted in H, even though, in fact, the error is occurring.) It is for this reason that many who were sympathetic to likelihood testing, including leaders of that approach, e.g., Hacking (1972), Barnard (1972), and Birnbaum (1969), gave it up.

17. Let l be the ratio of the likelihood of H and an alternative hypothesis J on given data x. We cannot say that because l is a small value that "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 [so small a likelihood ratio] in the case where the hypothesis tested [H] is true" (Pearson and Neyman, 1930, 106).

Thus, it seems that the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretation (Birnbaum, 1969, 128). Bayesians face the analogous problem, even with the addition of prior probabilities, as shown by Armitage (1962). (For a discussion, see Mayo, 1996, chap. 10; see also Mayo and Kruse, 2001; Mayo, 2003b.)

A test's error probabilities do matter, not because of an interest in having a good track record, but because of their role in informing us of the process that produced the data in front of us. We use them as standards to weed out erroneous interpretations of data. Those who insist that a genuine evidential appraisal is one of the E-R measures alone are demanding we renounce this use of error probabilities. This is something that scientists interested in interpreting data objectively are or should be unwilling to do.

CONCLUDING REMARKS

To summarize, the key difference between standard NP methods and E-R logics is that the former desires and is able to control error probabilities, whereas the latter is not. Criticisms of NP tests that are not merely misinterpretations stem from supposing that (1) the reason error probabilities matter in NP tests is solely from an interest in a low probability of erroneous "acts" in the long run and (2) an appropriate theory of evidence must provide an E-R measure. I have argued against both of these assumptions.

Error probabilities are used to assess the severity with which primary hypotheses of interest are tested by means of statistical evidence. By reporting on error probabilities (together with outcomes and sample sizes), the error statistician is providing information by which we can assess (at least approximately) whether data provide reliable evidence for experimental claims. Knowledge of the error-probabilistic properties of a test procedure—however that test was chosen—is the basis for criticizing a given test (e.g., as too sensitive or not sensitive enough) and for finding certain interpretations of evidence unwarranted. Nothing analogous can be said for criticizing subjective degrees of belief. Systematic rules may be specified to guide the appropriate interpretation of particular statistical results. Two such "metastatistical" rules that I propose are directed at interpreting rejections (rule of rejection) and acceptances (rule of acceptance). (See in Mayo, 1996, chap. 11. The reader will find considerable elaboration and updates of these ideas in chapter 4.3, my rejoinder to the commentaries on this chapter.)

The error-statistical account licenses claims about hypotheses that are

and are not indicated by tests without assigning quantitative measures of support or probability to those hypotheses. To those E-R theorists who insist that every uncertain inference have a quantity attached, our position is that this insistence is seriously at odds with the kind of inferences made every day, in science and in our daily lives. There is no assignment of probabilities (or other E-R measures) to the claims themselves when we say such things 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. What there is, instead, are arguments that a set of errors have (or have not) been well ruled out by appropriately severe tests.

4.1 Commentary

Earl D. McCoy

Scientists need philosophers to show them how science should operate. A deeper understanding of how evidence can be used to generate knowledge ultimately can provide scientists with more effective techniques for doing so. I would guess, however, that most scientists would be hard pressed to suggest how philosophical analysis actually has contributed to the incorporation of more effective techniques into their particular field of study. I also would guess that most scientists believe that philosophical analysis is the philosophers' game, with no real relevance to scientists' day-to-day practices. Of course, philosophers would retort that this position is borne out of a naïve viewpoint, and they are right. But if my guesses are correct, then what is to blame for these attitudes among most scientists? To be provocative, I will suggest that at least part of the blame must rest with philosophers. They sometimes seem to take positions on important issues that, even if philosophically consistent and well reasoned, appear to scientists to be out of touch with reality. I will suggest further, therefore, that just as scientists need philosophers, philosophers need scientists, to remind them of how science actually does operate. Deborah Mayo's chapter reinforces these two suggestions admirably.

Mayo first sets up the Bayesian model as the leading example of a logic of evidential relationship. She notes, however, the increasingly widespread

use of error-statistical methods in science. She then proceeds to set out a series of arguments to convince her readers to adopt the error-statistical (Neyman-Pearson, frequentist) model instead of the Bayesian (including the special-case maximum-likelihood) model. The principal advantage of the error-statistical model, she thinks, is that calculation of error probabilities is at its core. Indeed, whether intrinsic and supposedly unbiased error probabilities or extrinsic and often subjective probabilities based on experiences and beliefs are more relevant to the practice of science is the root of the debate between advocates of the two kinds of models. Mayo's arguments in favor of the error-statistical model are compelling. But two points strike me as strange. First, most scientists I know would not consider the Bayesian model to be the leading example of a logic of evidential relationship at all, but rather the error-statistical model. Second, while Mayo is encouraging a move away from the Bayesian model and toward the error-statistical model, increasing numbers of scientists are paying much closer attention to the Bayesian perspective on statistical analysis (see Malakoff, 1999). In my own discipline, ecology, several recent major examples of this trend can be noted (see, e.g., Hilborn and Mangel, 1997). What could lead to two such distinct, even contradictory, impressions of current scientific practice?

The answer to this question is not simple and may hinge on even deeper aspects of scientific inquiry. Mayo's arguments in favor of the error-statistical model appear to assume the central importance of the classic reductionist approach to scientific understanding. Many practicing scientists, on the other hand, are questioning the relevance of this approach to their research. Scientists who adopt this viewpoint often are those who need to incorporate social and environmental complexities into their analyses, rather than eliminate them; or who need to provide answers to questions that do not lend themselves readily to hypothesis testing; or who need to provide policy makers with realistic probabilities of uncommon but high-consequence events. In ecology, for example, practicing scientists often have suggested that the long time scales, lack of replication, and lack of controls inherent in many ecological studies prevent effective use of the classic reductionist approach. Many scientists, therefore, may view Mayo's arguments as irrelevant, even while agreeing with them in principle. In the best of circumstances, these scientists might advocate the error-statistical model. In the increasingly pervasive climate of uncertainty and urgency present in a variety of disciplines, however, the same scientists might see the Bayesian method (and other "unusual" methods, such as risk analysis and exploratory data analysis) as potentially useful.

Given this philosophical divide, the rational course may be to keep both the baby and the bath water. For one thing, the Bayesian model is not going to challenge the preeminence of the error-statistical model among scientists in the foreseeable future, because frequentist statistics have an established track record and Bayesian statistics do not. Furthermore, the two models may have different purposes: the error-statistical model may be the better approach for "basic" science (if such a thing really exists), but it may not be for "applied" science. The error-statistical model may not be the better approach for applied science because time may not permit us to ignore prior knowledge in solving applied problems and because the successful solution of many applied problems may not even require the development of full knowledge from the evidence at hand. I suggest that we simply admit that no way of knowing is perfect, keep both models, and use each of them appropriately and cautiously.

4.2 Commentary

George Casella

The thesis of Professor Mayo is that "error statistics," an approach that both assesses fit and attaches frequentist error probabilities to inferences, is the proper way for a statistician to produce an inference. Although I basically agree with this view, the framework that Mayo outlines may be too rigid to be applied in problems that the practicing statistician (data analyst, data miner!) faces today. However, we do need to apply these principles, but in a way that can actually be used in today's practice.

Mayo makes an eloquent argument in favor of the Neyman-Pearson (error statistics) school over the Bayesian (evidential relationship) school. But both approaches have their problems, and neither one is appropriate for all experiments. The Neyman-Pearson approach results in inferences that are valid against a long-run, objective assessment. In theory this is quite desirable, but in practice embedding an experiment in long-run trials doesn't always make sense. (For example, I recall an animal scientist whose data consisted of records of all dairy cattle in a particular country. As he noted, the experiment is not likely to be repeated.) Moreover, the classical Neyman-

This research was supported by NSF Grant No. DMS 9971586.

Pearson conclusion (accept or reject) is almost universally ignored by subject matter researchers in favor reporting the *P*-value.

The Bayesian (hierarchical) approach to modeling is well suited for complex problems, but the associated inference is consistently mired in its lack of objectivity. Although there are (seemingly countless) attempts at "objective" Bayes inference, even Bayesians cannot seem to agree on this, not to mention subject matter scientists.

But what do practicing statisticians need? How to proceed? Almost any reasonably complex model is a mixture of empirical and subjective components. When evaluating such models, the statistician must consider both frequency and Bayesian measures. Moreover, it may be the case that the empirical and subjective pieces are so intertwined as not to allow separate evaluations (for example, in empirical Bayes models the prior parameters are estimated from the data). Is such a model best evaluated by one or the other camp, or by a mixture? Mayo outlines a "framework of inquiry" that is compelling in its ideas, frustrating in its vagueness, and unworkable in its rigidity. However, the advice here is too good to disregard. Rather, it needs to be relaxed and expanded to better reflect what is happening in practice. Let us consider the three tasks of the framework of inquiry.

Task (i) instructs us to break large questions into small ones that can better be answered by data models. However, keep in mind that such low-level models are usually constructed after the data are seen, when helping an experimenter to make sense of the data. Thus, inferences will not be made using data collected from a simple model, but will be about a low-level model in the context of the entire experiment.

Task (ii) concerns the types of inferences to be made for the models of task (i). There is some good advice here, seemingly aimed at combating the accept/reject shortcomings of the Neyman-Pearson inference, but it is at a vague level and the possibility of its implementation is not clear. In particular, Mayo is describing an approach similar to Kiefer's conditional confidence (Kiefer, 1977). Despite a large amount of work on this topic (see, for example, Casella, 1988, or Goutis and Casella, 1995), it is not yet ready for prime time. Work by Berger, Brown, and Wolpert (1994) and Berger, Boukai, and Wang (1997) has promise, but is still confined to more stylized problems than one meets in practice.

Task (iii) seems to say, "do a power calculation first" and, after the data have been gathered, "check residuals." In general, this is very sound advice.

What puzzles me is that I cannot see why the three tasks cannot be accomplished within a Bayesian framework too (not instead). The low-level modeling of task (i) can be accomplished equally well within the Bayesian paradigm, and the power calculations of task (iii) are also straightforward for a Bayesian, who will usually agree to average over the sample space before the data are seen. This leaves us with how to assess error probabilities and fit. But this can be accomplished by looking at posterior probabilities and credible regions, along with a sensitivity analysis. If one accepts that the Bayes inference is based on a subjective premise, and the effect of the premise can be assessed through a sensitivity analysis, then the error assessment of task (ii) can be done equally well within the Bayesian framework. Then, using both Neyman-Pearson and Bayesian tools, we can apply these very sound principles to the type of complex models that are being used in practice.

4.3 Rejoinder

Deborah G. Mayo

I am grateful to professors McCoy and Casella for their interesting and enlightening comments on my paper, and grateful too for this chance to clear up some misunderstandings and lack of clarity as to what I am advocating. My response will discuss: (a) flawed philosophical assumptions about objectivity in science, (b) problems with what might be called the "conditional error probability (CEP)" movement, and (c) an illustration of the severity reinterpretation of NP tests.

THE "OLD" (LOGICAL POSITIVIST) IMAGE OF SCIENCE: IT IS TIME TO REJECT THE FALSE DILEMMA

To begin with, I find the remarks of both these scientists to be premised on a false dilemma based on a discredited philosophy of science and of scientific objectivity, and in this first part of my comments, I will attempt to explain (1) how to escape the horns of their dilemma, and (2) why doing so is so important precisely for the practicing scientists with whom both claim to be concerned. The false dilemma stems from the supposition that either scientific inference is "reducible" to purely objective, value-free methods or else they are ineluctably subjective, relative, and more a matter of belief and opinion than of strictly empirical considerations. When this false dilemma is coupled with the assumption that the "error-statistical" philosophy of sta-

tistics goes hand in hand with the former ("reductionist") position, it is easy to see why McCoy and Casella argue from the impossibility of a value-free science to conclude that scientists must embrace a subjective-relativist position, as reflected, for instance, in subjective Bayesian statistics. For example, McCoy alleges that my "arguments in favor of the error-statistical model appear to assume . . . the classic reductionist approach" to science, in contrast to practicing scientists who, recognizing the "need to incorporate social and environmental complexities into their analyses" are led to use Bayesian methods, wherein such complexities may find cover under subjective degrees of belief. Casella too argues that since in scientific practice "it may be the case that the empirical and subjective pieces are so intertwined as not to allow separate evaluations," it follows that scientists need to use both "frequency and Bayesian measures." Both arguments are fallacious.

Granted, logical positivist philosophers had long encouraged the traditional or "old image" of a purely objective, value-free science that underwrites McCoy and Casella's remarks, but the philosophies of science and statistics I favor are based on denying both disjuncts that they appear to regard as exhaustive of our choices. As such, these charges have no relevance to my position. As I make clear in my paper, I deny the classic logical positivist position that a genuinely objective scientific inference must be reducible to a logic or algorithm applied to (unproblematically) given statements of data and hypotheses. At the same time I deny the assumption of many a postpositivist (led by Kuhn and others) that the untenability of any such algorithm or "logic of evidential-relationship" (E-R logic) leaves us with no choice but to embrace one of the various stripes of relativistic or subjectivistic positions, as found both in the context of general philosophy of knowledge (social-constructivism, psychologism, postmodernism, anarchism) and in philosophy of statistics (e.g., subjective Bayesianism).

We can readily agree with Casella that "any reasonably complex model is a mixture of empirical and subjective components" while adamantly denying his inference that therefore "the statistician must consider both frequency and Bayesian measures." It is a mistake, a serious one, to suppose that the empirical judgments required to arrive at and evaluate complex models are no more open to objective scrutiny than are subjective Bayesian degrees of belief. The former judgments, in my view, are not even captured by prior probabilities in hypotheses, and they are open to criticism in ways that subjective degrees of belief are not. For McCoy to imply, as he does, that the error-statistical model requires us "to ignore prior knowledge" is to misunderstand the crucial ways in which background knowledge must be used in each of the tasks of collecting, modeling, and interpreting data. Summaries of subjective degrees of belief may seem to offer a more formal way of importing information into inference, but this way fails to serve the error statistician's goals.

Moreover, from the recognition that background beliefs and values may enter at every stage of scientific inference, as I have long argued (e.g., Mayo, 1989, 1991, 1996), we can develop explicit methods for critically evaluating results of standard statistical tests. Such a metastatistical scrutiny enables us to discern what the data do and do not say about a phenomenon of interest, often in terms of the existence and direction of a discrepancy from a hypothesized value (I take this up below). I am unaware of an existing subjective Bayesian approach that achieves these ends or one that promotes the critical scrutiny needed to understand and adjudicate conflicting interpretations of evidence.

Once again, I could not agree more with Casella and McCoy in their demand that any adequate account be able to deal with what Casella calls the "problems that the practicing statistician (data analyst, data miner) faces today," and to grapple with the "lack of replication, and lack of controls inherent in many ecological studies" that McCoy refers to (on data mining, see Spanos, 2000). However, it is precisely because of the complexities and limitations in actual experimental practice that scientists are in need not of a way to quantify their subjective beliefs as Casella and McCoy recommend (much less to allow their opinions to color the inference as they do in a posterior degree-of-belief calculation), but rather of a way to check whether they are being *misled* by beliefs and biases, in scrutinizing both their own data and that of other researchers. An adequate account of scientific inference must provide a methodology that permits the consequences of these uncertainties and knowledge gaps to be criticized by others, not permit them to be kept buried within a mixture of subjective beliefs and prior opinions.

In this connection, I must protest McCoy's allegation that scientists faced with "the increasingly pervasive climate of uncertainty and urgency" of practice may view my arguments as "irrelevant." The truth is that every step of my error-statistical (re)interpretation of standard (Neyman-Pearson) statistical methods grew directly out of the real and pressing concerns that arise in applying and interpreting these methods, especially in the face of the "uncertainty and urgency" of science-based policy regarding risky technologies and practices. (See, for instance, my sketch of the "rule of acceptance" below). Where McCoy touts the subjective Bayesian model because it licenses scientists in providing policy makers with probabilities of "uncommon but high-consequence events," for many of us, the increasing dependence on such "expert" assessments that cannot be readily critiqued, and

which too often have been seriously wrong, is more a cause for alarm than it is a point in favor of the Bayesian model. (To my mind, the very reason that a philosopher such as myself was invited to the conference upon which this volume is based is precisely the concern engendered by the increased tendency of policy makers to rely on Bayesians who are only too happy to give us "expert" assessments of such things as how improbable it is for a given species to go extinct, or for a given transgenic food to adversely affect untargeted species, and so on, never mind the uncertainties, the effects not yet even tested, etc.)

As I see it, the appropriateness of tools for responsible scientific practice should be measured not according to whether they can cough up the kinds of numbers that satisfy policy makers—who may not be fully aware of the uncertainties and complexities involved—but according to how good a job they do in interpreting and communicating the nature of the uncertainties and errors that have and have not been well ruled out by existing evidence (e.g., "there is evidence from the lab that a genetically modified crop does not harm untargeted species, but we have yet to test it in the field"). While this may result in far less tidy reports than some might wish, this may be the only responsible scientific report that can be given, and it should be left to the public and to policy makers to decide what management decisions are appropriate in the face of existing uncertainties (e.g., taking precautions).

In practice, adopting the premise that McCoy and Casella blithely assume—that unless statistical inference is value-free we are forced to embrace subjectivism—exacerbates the problem of holding risk managers accountable to the public who are affected most by science-based policies that are made in the "climate of uncertainty and urgency." If statistical inferences on which risk judgments are based are really as inextricably entwined with social policy opinions as my commentators appear to assume, there would be little grounds for criticizing a risk assessment as a misinterpretation of data, in the sense of incorrectly asserting what the data indicate about the actual extent of a given risk. Statistical methods become viewed as largely tools for manipulation rather than as instruments to help achieve an unbiased adjudication of conflicting assessments. In the current climate of widespread public suspicion that "scientific expertise" can be bought in support of one's preferred policy, what is needed most is neither tidy degree-of-belief judgments nor despairing evidence-based risk assessments. What are needed are principled grounds to solicit sufficient information to enable statistical reports to be criticized and extended by others. This requires, in the realm of statistical reports, information about the *error probabilities* of

methods—the very thing that is absent from subjective Bayesian accounts. This leads to my second set of remarks.

THE THREE TASKS: WHY BAYESIANS (AND OTHER CONDITIONALISTS) ARE WRONG FOR THE JOB

I am struck by Casella's suggestion that he "cannot see why the three tasks" that I list for applying error-statistical methods to scientific inquiry "cannot be accomplished within a Bayesian framework" as well. Perhaps they can, but it is important to see that they will no longer be the same tasks as those understood by the error statistician. I will focus on task (ii) (carrying out error-statistical tests), but a few words about tasks (i) and (iii) are in order.

A central obstacle to carrying out the kind of piecemeal analysis of task (i) within a Bayesian account is that Bayes' theorem requires one to work with "a single probability pie" (Mayo, 1997a, 231), as it were. One cannot escape the assessment of the Bayesian "catchall factor," $P(x|\text{not-}H)$, in making inferences about H, where not-H will include all hypotheses other than H, including those not yet even dreamt of. The error-statistical assessment, by contrast, permits probing local hypotheses without having to consider, much less assign probabilities to, an exhaustive space of alternative hypotheses.

When it comes to testing the assumptions of statistical test data [task (iii)] the error statistician has a principled justification for appealing to pretrial methods, e.g., randomization, to generate data satisfying error-statistical assumptions, as well as a methodology for running posttrial tests of assumptions (e.g., independence and identical distribution). Once the data $\boldsymbol{X}$ are in hand, the likelihood principle entails the irrelevance of the procedures generating $\boldsymbol{X}$ (e.g., stopping rules, randomization) that do not alter likelihoods even though they can dramatically alter error probabilities (Mayo, 1996, 2003a; Mayo and Kruse, 2001). Moreover, if probability models reflect subjective degrees of belief (or ignorance), how are they to be put to a test? What can it even mean for them to be "wrong" (Lindley, 1976)? Even non-subjective Bayesians are limited to appeals to "robustness" and asymptotics that come up short.

I now turn to focus in some detail on task (ii), carrying out statistical tests in a manner that avoids the shortcomings of the rigid NP model.

THE NEW CONDITIONAL ERROR PROBABILITY (CEP) MOVEMENT

Let me readily grant Cassella's charge that my proposals in chapter 4 remained at "a vague level"—something I rectify somewhat in part (c) of these comments. As for Casella's statement that I am "describing an approach similar to Kiefer's conditional confidence (Kiefer, 1977)," I cannot say (though I hope to probe this further). However, I am less sanguine that the goals of my reinterpretation could be met by the contemporary "conditional error probability" (CEP) movement to which Cassella also alludes, e.g., as promoted by Jim Berger, even if extended beyond it's current "stylized" examples.

As intriguing as I find these Bayesian-error-statistical links, the notion of "conditional error probabilities" that arises in this work refers not to error probabilities in the error statistician's sense but rather to posterior probabilities in hypotheses. These differ markedly from NP error probabilities, which refer only to the distribution of test statistic $d(\boldsymbol{X})$, i.e., the *sampling distribution.* True, my use of the sampling distribution (for a post-data evaluation of severity) goes beyond what is officially prescribed in the NP model; nevertheless, all of my probabilistic assessments are in terms of the distribution of $d(\boldsymbol{X})$, and do not involve any posterior probability assignment. Hence, it is misleading to refer to the conditionalist's posterior probability assignments as "conditional error probabilities"; to do so is sure to encourage, rather than discourage, the confusion that already abounds in interpreting error probabilities such as observed significance levels or *P*-values.

Calling these posterior probabilities in H "error probabilities" or even "conditional error probabilities" is misleading. The posterior probabilities of H, arrived at via one of these conditional approaches (e.g., by stipulating an "uninformative prior") will not provide control of error probabilities as understood in the NP testing account. Accordingly, they will not underwrite the severity assessments that are the linchpins of my approach. Ironically, however, CEPs are often billed as "correcting" frequentist error probabilities by introducing priors that are allegedly kosher even for a frequentist.

Consider a very common illustration. In a normal distribution (two-sided) test of H_0: $\mu = 0$ versus H_1: $\mu \neq 0$, "at least 22%—and typically over 50%—of the corresponding null hypotheses will be true" if we assume that "half of the null hypotheses are initially true," conditional on a .05 statistically significant result (Berger, 2003). This is taken to show the danger in interpreting *P*-values as error probabilities, but note the shift in meaning of "error probability." The assumption is that the correct error probability is given by the proportion of true null hypotheses (in a chosen population of nulls). But an error statistician would (or should) disagree.

Take the recent studies reporting statistically significant increases in breast cancer among women using hormone replacement therapy HRT for 5 years or more. Let us suppose P is .02. (In a two-sided test this is approximately a .05 *P*-value. Because the 1-sided test has much less of a discrepancy between *P*-values and posteriors, the critics focus on the 2-sided, and I do as well.)

The probability of observing so large an increase in disease rates when H_0 is true and HRT poses no increased risk is small (.05). Given that the assumptions of the statistical model are met, the error-statistical tester takes this to indicate a genuine increased risk (e.g., 2 additional cases of the disease per 10,000). To follow the CEP recommendations, it would seem, we must go on to consider a pool of null hypotheses from which H_0 may be seen to belong, and consider the proportion of these that have been found to be true in the past. This serves as the prior probability for H_0. We are then to imagine repeating the current significance test p over all of the hypotheses in the pool we have chosen, and look to the posterior probability of H_0 to determine whether the original assessment of a genuine increased risk is or is not misleading. But which pool of hypotheses should we use? Shall we look at all those asserting no increased risk or benefit of any sort? Or only of cancer? In men and women? Or women only? Of breast cancer or other related cancers? Moreover, it is hard to see how we could ever know the proportion of nulls that are true rather than merely those that have thus far not been rejected by statistical tests.

Finally, even if we agreed that there was a 50% chance of randomly selecting a true null hypothesis from a given pool of nulls, that would still not give the error statistician's frequentist prior probability of the truth of the given null hypothesis, e.g., that HRT has no effect on breast cancer risks. Either HRT increases cancer risk or not. Conceivably, the relevant parameter, say the increased risk of breast cancer, could be modeled as a random variable (perhaps reflecting the different effects of HRT in different women), but its distribution would not be given by computing the rates of other benign or useless drugs!

These allegedly frequentist priors commit what might be called the fallacy of instantiating probabilities:

> p% of the null hypotheses in a given pool of nulls are true.
> This particular null hypothesis H_0 was randomly selected from this pool.
> Therefore $P(H_0 \text{ is true}) = p$.

Admittedly, most advocates of Bayesian CEP, at least if pressed, do not advocate looking for such a frequentist prior (though examples abound in the literature), but instead advocate accepting from the start the "objective" Bayesian prior of .5 to the null, the remaining .5 probability being spread out over the alternative parameter space. But seeing how much this would influence the Bayesian CEP and how this in turn would license discounting the evidence of increased risk should make us that much more leery of assuming them from the start. Now, the Bayesian significance tester wishes to start with a fairly high prior to the null—the worry being that otherwise a rejection of the null would amount to reporting that a fairly improbable hypothesis has become even more improbable. However, it greatly biases the final assessment toward finding no discrepancy. In fact, with reasonably large sample size n, a statistically significant result leads to a posterior probability in the null that is higher than the initial "ignorance" prior of .5 (see, e.g., Berger, 2003).

Hence, what the error statistician would regard as a good indication of a positive discrepancy from 0 would, on this conditional approach, not be taken as any evidence against the null at all. While conditionalists view this as rectifying the alleged unfairness toward the null in a traditional significance test, for the error statistician this would conflict with the use of error probabilities as postdata measures of the *severity* of a test. What the severity tester would regard as good evidence for a positive discrepancy from H_0 is no longer so regarded because the evidence of a discrepancy has been washed away by the (high) prior probability assignment to H_0 (Mayo, 2003b).

Note the consequences of adopting such an approach in the "pervasive climate of uncertainty" about which McCoy is concerned. Uncertainty about a positive risk increase due to an unfamiliar technology would seem to license the "ignorance prior" probability of .5 to hypothesis H_0: 0 risk increase. Accepting the conditionalist strategy exacerbates the problem that already exists in finding evidence of risk increases with low-probability events, except that there would be no grounds for criticizing the negative report of low confirmation of risk.

The severity assessment *is* applicable to those cases where a legitimate frequentist prior probability distribution exists,[18] and the aim is inferring claims about posterior probabilities. In such cases, the severity assessment would be directed at the reliability of these probability assignments and at

18. Editors' note: see Goodman, 2004 (chapter 12 of this volume), for another discussion of the use of empirical priors.

checking for mistakes in the various ingredients of the Bayesian calculation involved.

Given that I find myself so often seconding Casella's own criticisms of (CEP) attempts (e.g., Cassella and Berger, 1987), I remain perplexed that he seems to regard at least the broad line of the CEP movement as a promising way to resolve problems in the philosophy of statistics. Why not prefer, instead, a resolution that does not entail renouncing frequentist principles and philosophy?

To be fair, Casella appears to harbor an assumption that is nearly ubiquitous among contemporary critics of frequentist methods: namely that the way to supply NP tests with a viable postdata inferential interpretation is by developing strategies wherein NP error probabilities coincide with postdata probabilities on hypotheses.

The underlying assumption is that for error probabilities to be useful for inference they need to provide postdata measures of confirmation to hypotheses—i.e., E-R measures—and that such measures are afforded by posterior probability assignments to hypotheses. The central thesis of my chapter, however, is to reject this assumption. My postdata interpretation, in contrast to that of the conditionalist, uses NP error probabilities to assess the severity with which a particular inference passes a test with data x; and it is a mistake to construe a postdata severity assessment as a measure of support, probability, or any other E-R measure to the statistical hypotheses involved!

THE SEVERITY REINTERPRETATION OF STATISTICAL HYPOTHESIS TESTS

Now to go at least part way towards addressing Casella's charge that my reinterpretation of NP methods remains "at a vague level." The quickest way to put some flesh on the ideas of my paper, I think, will be by means of a very familiar statistical test—the one-sided version of the test we just considered.

Example: Test T_α. Suppose that a random sample of size n, $\boldsymbol{X} = (X_1, \ldots, X_n)$, where each X_i is normally distributed with unknown mean μ and known standard deviation $\sigma = 1$, is used to conduct a *one-sided test* of the hypotheses

H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$.

The NP test is a rule that tells us for each possible outcome $\boldsymbol{x} = (x_1, \ldots, x_n)$ whether to "accept H_0" or "reject H_0 and accept H_1." The rule is defined in terms of a test statistic or distance measure $d(\boldsymbol{X})$

$$d(\boldsymbol{X}) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

where $\overline{X}$ is the sample mean. In particular, the uniformly most powerful (UMP) test with significance level α, denoted by T_α, is

T_α: reject H_0, if $d(x) \geq c_\alpha$.

Equivalently,

T_α: reject H_0, if $\overline{X} \geq \mu_0 + c_\alpha \sigma_x$,

where $\sigma_X = \dfrac{\sigma}{\sqrt{n}}$. Setting $\alpha = .02$ corresponds to a c_α of approximately 2, and so $d(\boldsymbol{X})$ would be *statistically significant* (at the .02 level) whenever $d(\boldsymbol{X}) > 2$.

To have a specific numerical example, let $n = 100$, so $\sigma_x = 0.1$ (since we are given that $\sigma = 1$), and we have

$T_{.02}$: reject H_0 whenever $d(\boldsymbol{X}) \geq 2$.

THE FORM OF THE REINTERPRETATION

Although NP tests are framed in terms of a hypothesis being "rejected" or "accepted," both results will correspond to "passing" some hypothesis or, rather, some hypothesized discrepancy, enabling a single notion of severity to cover both. That is, we use the observed $\boldsymbol{x}$ that leads to "reject" or "accept" H_0 in order to assess the *actual severity* with which specific discrepancies from H_0 pass the given test. "Reject H_0" in this one-sided test will license inferences about the extent of the positive discrepancy that is *indicated* by data $\boldsymbol{x}$; whereas "accept H_0" will correspond to inferences about the extent of the discrepancy from H_0 that is *ruled out:*

Reject H_0 (with $\boldsymbol{x}$) licenses inferences of the form $\mu > \mu'$.
Accept H_0 (with $\boldsymbol{x}$) licenses inferences of the form $\mu \leq \mu'$.

It is very important to emphasize that this is *not* to change the null and alternative hypotheses associated with test T_α; they are the same ones given at the outset. Rather, these are postdata inferences that might be considered in interpreting the results of running T_α. The particular inference that is *warranted* will depend on the corresponding *severity* assessment. This brings out a crucial contrast between a severity assessment and the probabilities of type I and type II errors (or power), namely that a severity assessment is always relative to a particular hypothesis or inference one is entertaining. The question "Is T_α a severe test?" is not well posed until one specifies the particular inference being entertained and with what test result. This will become clearer as we proceed.

Severity

To implement the reinterpretation, we need to remind ourselves of the definition of severity. The basic idea of a hypothesis H passing a severe test with data $\boldsymbol{x}$ is the requirement that it be highly improbable that H would have passed so well if in fact H is false (e.g., if a specified discrepancy from H exists). More precisely, a hypothesis H passes a severe test with data $\boldsymbol{x}$ if (and only if),

(i) $\boldsymbol{x}$ agrees with H (for a suitable measure of agreement or fit), and
(ii) with very high probability, the test would have produced an outcome that fits H *less well* than $\boldsymbol{x}$ does if H were false or incorrect.

Alternatively, we can write clause (ii) as

(ii′) there is a very *low* probability that the test would have produced an outcome that fits H *as well* as $\boldsymbol{x}$ does, if H were false or incorrect (Mayo, 1996).

The basic tenet underlying the interpretation of test results is that inferences are warranted just to the extent that they have passed severe tests. We can call this the severity principle:

Severity principle: Data $\boldsymbol{x}$ provide a good indication of, or good evidence for, hypothesis H to the extent that H passes a severe test with $\boldsymbol{x}$.

Rules of Rejection and Acceptance: The postdata interpretation for the results of test T_α may be expressed in terms of two rules associated with "reject H_0" and "accept H_0" respectively.

Rules of Rejection (RR) for T_α (i.e., $d(\boldsymbol{x}) > c_\alpha$):

(a) If there is a very *low* probability of obtaining so large a $d(\boldsymbol{X})$ even if $\mu \leq \mu'$ then $d(x)$ passes hypothesis $\mu > \mu'$ with *high* severity; hence, by the *severity principle*, $d(x)$ provides *good evidence* for $\mu > \mu'$.

The severity principle also tells us which discrepancies we would *not* be warranted to infer, and we can state this in a companion rule:

(b) If there is a very *high* probability of obtaining so large a $d(x)$ even if $\mu \leq \mu'$ then $d(x)$ passes hypothesis $\mu > \mu'$ with *low* severity; hence, by the *severity principle*, $d(x)$ provides *poor evidence* for $\mu > \mu'$.

Analogously, we can state rules for interpreting "accept H_0", this time combining both parts (a) and (b):

Rules of Acceptance (RA) for test T_α (i.e., $d(x) \leq c_\alpha$:—a "negative" result):

(a) and (b): If there is a very *high* (*low*) probability that $d(x)$ would have been larger than it is, (even) if $\mu > \mu'$, then $d(x)$ passes hypothesis $\mu \leq \mu'$ with *high* (*low*) severity; hence, by the severity principle, $d(x)$ provides *good (poor) evidence* for $\mu \leq \mu'$.

Defining Severity for T_α: The severity with which hypothesis H passes T_α with an observed $d(x_0)$, abbreviated as $\text{Sev}(T_\alpha; H, d(x_0))$, where x_0 is the actual outcome:

The Case of Rejecting H_0 (i.e., $d(x_0) > c_\alpha$):
$\text{Sev}(T_\alpha; \mu > \mu', d(x_0)) = P(d(\boldsymbol{X}) \leq d(x_0); \mu > \mu' \text{ is false})$.

The Case of Accepting H_0 (i.e., $d(x_0) < c_\alpha$):
$\text{Sev}(T_\alpha; \mu \leq \mu', d(x_0)) = P(d(\boldsymbol{X}) > d(x_0); \mu \leq \mu' \text{ is false})$.

In both cases, we calculate severity under the point value μ' giving the lower bounds for severity.[19]

PARTICULAR APPLICATIONS AND REMARKS

To get a feel for applying these rules using the definition above, consider some particular outcomes from test $T_{.02}$ with $n = 100$. Recall that

19. Editors' note: similar to the power curve, this statement actually defines a curve, not a single probability number, as μ' varies.

$T_{.02}$: reject H_0 whenever $d(x) \geq 2$ (rounding for simplicity).

Suppose we are testing

H_0: $\mu = 0$ against H_1: $\mu > 0$.

Consider first some examples for applying RR: Let $\bar{x} = 0.2$—an outcome just at the cutoff for rejecting H_0. We have $P(\bar{X} \leq 0.2; \mu = 0) = P(Z \leq 2) = .977$, and thus from rule RR we can infer with severity .98 that $\mu > 0$. Further, because $P(\bar{X} \leq 0.2; \mu = .1) = P(Z \leq 1) = .84$ we can infer with severity .84 that $\mu > 0.1$.

Note in general for tests of type T_α, that by stipulating that H_0 be rejected only if $d(\boldsymbol{X})$ is statistically significant at some small level α, it is assured that such a rejection—*at minimum*—warrants hypothesis H_1, that $\mu > \mu_0$. Rule RR(a) makes this plain.

Severity vs. Power

At first glance, severity requirement (ii) may seem to be captured by the notion of a test's power. However, the severity with which alternative $\mu > \mu'$ passes a test is not given by, and is in fact inversely related to, the test's power to reject H_0 in favor of μ'. For instance, the severity with which $\bar{x} = 0.2$ passes hypothesis H: $\mu > 0.4$ is .03, so it is a very poor indication that the discrepancy from 0 is this large. However, the power of the test against 0.4 is high, .97 (it is in the case of "accept H_0" that power correlates with severity, as seen below).

Beyond the Coarse Interpretation of the NP Test

If $d(\boldsymbol{X})$ exceeds the cut-off 2, this is reflected in the severity assessments. For example, if $\bar{x} = 0.3$, we have from rule RR:

infer with severity .999 that $\mu > 0$, and also
infer with severity .98 that $\mu > 0.1$.

Consider now the case where test T_α outputs "accept H_0." We know that failing to reject H_0 does not license the inference to H_0—that μ is *exactly* 0—because the probability of such a negative result is high even if there is *some* positive discrepancy from 0. Hence, the well-known admonition that "no evidence against (the null) is not the same as evidence for (the null)!" Rule RA explains this in terms of severity. However, RA may also be used to find a μ' value that *can* be well ruled out.

Suppose $\bar{x} = 0.1$. $P(\bar{X} > 0.1; \mu = 0.05) = P(Z > 0.5) = 0.31$, and from rule RA we can infer with severity (only) .31 that $\mu < 0.05$. But also, because $P(\bar{X} > 0.1; \mu = 0.2) = P(Z > -1) = .84$, we can infer with severity .84 that $\mu < 0.2$.

Thus, imagine on the basis of $\bar{x} = 0.1$, someone proposes to infer "This result is evidence that if there are any discrepancies from 0, it is less than 0.05." Since the severity with which $\mu < 0.05$ passes this test is only 0.3, we would criticize such an interpretation as unwarranted—and we would do so on objective grounds. One might go on to note that the evidence can only be regarded as passing a much less informative claim, i.e., $\mu < 0.2$, with reasonable severity (e.g., .84). By making this type of critique systematic, the severity assessment supplies a powerful tool for critically evaluating negative reports, especially in cases of statistical risk assessment (e.g., Mayo, 1985b, 1989, 1991, 1996). Using severity to interpret nonsignificant results thereby supplies a data dependency that is missing from recent attempts at "postdata power analyses" (Mayo and Spanos, 2000, 2004).

Answering Central Criticisms of NP tests

This postdata reinterpretation gives the following response to the alleged arbitrariness in choosing the probabilities for type I and II errors. Predata, the choices should reflect the goal of ensuring the test is capable of licensing given inferences severely. We set the "worst case" values accordingly: small α ensures, minimally, that any rejection of H_0 licenses inferring *some* discrepancy from H_0; and high power against μ' ensures that *failing* to reject H_0 warrants inferring $\mu \leq \mu'$ with high severity. Postdata, the severity assessment allows for an objective interpretation of the results of whatever test happened to be chosen. While statistical tests cannot themselves tell us which discrepancies are of substantive importance, the concept of severity lets us use the test's properties to assess both the discrepancies that have, and that have not, been passed severely. This is a powerful source for critically evaluating statistical results—something sorely needed in a great many applications of tests.

Summary of the Rules

We can summarize rules RR and RA for the case of test T_α with result $d(x_0)$ as follows:

RR: reject H_0 (i.e., $d(x_0) > c_\alpha$)

infer with severity $1 - \varepsilon$: $\mu > \bar{X} - k_\varepsilon\sigma/\sqrt{n}$, where $P(Z > k_\varepsilon) = \varepsilon$

or

infer with severity $1 - \varepsilon$: $\mu > \mu_0 + \gamma$, where $\gamma = (d(x_0) - k_\varepsilon)\sigma/\sqrt{n}$.[20]

RA: accept H_0 (i.e., $d(x_0)) \leq c_\alpha$):

infer with severity $1 - \varepsilon$: $\mu \leq \bar{X} - k_\varepsilon\sigma/\sqrt{n}$, where $P(Z > k_\varepsilon) = \varepsilon$

or

infer with severity $1 - \varepsilon$: $\mu \leq \mu_0 + \gamma$, where $\gamma = (d(x_0) + k_\varepsilon)\sigma/\sqrt{n}$.

In this form, there is a clear connection with NP confidence intervals, often recommended for a postdata inference. There are important differences, however, most notably that confidence intervals (in addition to having a problematic postdata interpretation) treat all values in the interval on par, whereas each value in the interval corresponds to inferences (about discrepancies from that value) with different degrees of severity. A full comparison between a severity assessment and confidence intervals calls for a separate discussion (Mayo and Spanos, 2000).

Also beyond the scope of these comments is a full discussion of how to use (and avoid misusing) the admittedly unfamiliar expression *infer with severity* $1 - \varepsilon$ *that* $\mu > \mu'$ $[\mu_0 + \gamma]$. One warning must suffice: $1 - \varepsilon$ should *not* be construed as a degree of support or other E-R measure being accorded to $\mu > \mu'$. It is just a shorthand for the longer claim: hypothesis $\mu > \mu'$ passes test T with severity $1 - \varepsilon$.

Examples of the kinds of questions that this reinterpretation allows one to pose are: *How severely does a given hypothesis pass T with a given result?* and *Which alternatives to* H_0 *pass T at a given level of severity?* Other types of questions can also be underwritten with this postdata interpretation, and it would be a matter for practitioners to consider which are most useful for given applications.

REFERENCES

Armitage, P. 1962. Contribution to Discussion. In Savage, L. J., ed., *The Foundations of Statistical Inference: A Discussion.* London: Methuen.

Barnard, G. A. 1962. Contribution to discussion. In Savage, L. J., ed., *The Foundations of Statistical Inference: A Discussion.* London: Methuen.

20. Here Z is the standard normal variate.

Barnard, G. A. 1972. Review of *The Logic of Statistical Inference*, by I. Hacking). *B. J. Phil. Sci.* 23:123–132.

Berger, J. O. 2003. Could Fisher, Jeffreys, and Neyman Have Agreed on Testing? *Stat. Sci.* 18:1–12.

Berger, J. O., B. Boukai, and Y. Wang. 1997. Unified Frequentist and Bayesian Testing of a Precise Hypothesis (with discussion). *Stat. Sci.* 12:133–160.

Berger, J. O., L. D. Brown, and R. L. Wolpert. 1994. A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing. *Ann. Stat.* 22:1787–1807.

Berger, J. O., and Wolpert, R. L. 1988. *The Likelihood Principle.* 2nd ed. Hayward, CA: Institute of Mathematical Statistics.

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

Birnbaum, A. 1977. The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; with a Criticism of the Lindley-Savage Argument for Bayesian Theory. *Synthese* 36:19–49.

Carnap, R. 1962. *Logical Foundations of Probability.* Chicago: University of Chicago Press.

Casella, G. 1988. Conditionally Acceptable Frequentist Solutions (with discussion). In Gupta, S. S., and J. O. Berger, eds. *Statistical Decision Theory IV.* New York: Springer.

Casella, G., and R. L. Berger. 1987. Commentary of J. O. Berger and M. Delampady's "Testing Precise Hypotheses." *Stat. Sci.* 2:344–347.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics.* London: Chapman and Hall.

de Finetti, B. 1972. *Probability, Induction, and Statistics: The Art of Guessing.* New York: Wiley.

Edwards, W., H. Lindman, and L. Savage. 1963. Bayesian Statistical Inference for Psychological Research. *Psych. Rev.* 70:193–242.

Gigerenzer, G. 1993. The Superego, the Ego, and the Id in Statistical Reasoning. In Keren, G. and C. Lewis, eds., *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.* Hillsdale, NJ: Erlbaum.

Goutis, C. and Casella, G. 1995. Frequentist Post-data Inference. *Internat. Stat. Rev.* 63:325–344.

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

Hacking, I. 1972. Likelihood. *Br. J. Phil. Sci.* 23:132–137.

Hilborn, R., and M. Mangel. 1997. *The Ecological Detective: Confronting Models with Data.* Princeton: Princeton University Press.

Howson, C., and P. Urbach. 1993. *Scientific Reasoning: The Bayesian Approach.* 2nd ed. La Salle, IL: Open Court.

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

Kiefer, J. 1977. Conditional Confidence Statements and Confidence Estimators (with discussion). *J. Am. Stat. Assn* 72:789–827.

Lindley, D. V. 1976. Bayesian Statistics. In Harper, W. L., and C. A. Hooker, eds., *Foundations and Philosophy of Statistical Inference.* Dordrecht: Reidel.

Malakoff, D. 1999. Bayes Offers a "New" Way to Make Sense of Numbers. *Science* 286:1460–1464.

Mayo, D. G. 1983. An Objective Theory of Statistical Testing. *Synthese* 57 (pt. 2): 297–340.

Mayo, D. G. 1985a. Behavioristic, Evidentialist, and Learning Models of Statistical Testing. *Phil. Sci.* 52:493–516.

Mayo, D. G. 1985b. Increasing Public Participation in Controversies Involving Hazards: The Values of Metastatistical Rules. *Sci. Tech. Human Values* 10:55–68.

Mayo, D. G. 1989. Toward a More Objective Understanding of the Evidence of Carcinogenic Risk. *PSA: Proceedings of the BiennialMeeting of the Philosophy of Science Association,* 1988, vol. 2: 489–503.

Mayo, D. G. 1990. Did Pearson Reject the Neyman-Pearson Philosophy of Statistics? *Synthese* 90:233–262.

Mayo, D. G. 1991. Sociological versus Metascientific Views of Risk Assessment. In Mayo, D., and R. Hollander, eds., *Acceptable Evidence: Science and Values in Risk Management.* New York: Oxford University Press.

Mayo, D. G. 1996. *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.

Mayo, D. G. 1997a. Duhem's Problem, The Bayesian Way, and Error Statistics, or "What's Belief Got to Do with It?"; and response to Howson and Laudan. *Phil. Sci.* 64:222–244, 323–333.

Mayo, D. G. 1997b. Error Statistics and Learning from Error: Making a Virtue of Necessity. *Phil. Sci.* 64 (supp.): S195–S212.

Mayo, D. G. 2003a. Severe Testing as a Guide for Inductive Learning. In Kyburg, H., and M. Thalos, eds., *Probability Is the Very Guide in Life.* Chicago: Open Court.

Mayo, D. G. 2003b. Commentary on J. O. Berger's Fisher Address ("Could Fisher, Jeffreys, and Neyman Have Agreed?"). *Stat. Sci.* 18:19–24.

Mayo, D. G. 2004. Evidence as Passing Severe Tests: Highly Probable vs. Highly Probed Hypotheses. In Achinstein, P., ed., *Scientific Evidence: Philosophical Theories and Applications.* Baltimore: Johns Hopkins University Press.

Mayo, D. G., and M. Kruse. 2001. Principles of Inference and Their Consequences. In Cornfield, D., and J. Williams, eds., *Foundations of Bayesianism.* Dordrecht: Kluwer.

Mayo, D. G., and A. Spanos. 2000. *A Post-data Interpretation of Neyman-Pearson Methods Based on a Conception of Severe Testing.* London: Tymes Court.

Mayo, D. G., and A. Spanos. 2004. Methodology in Practice: Statistical Misspecification Testing. *Phil. Sci.* (vol. 2 of PSA 2002 proceedings).

Neyman, J. 1955. The Problem of Inductive Inference. Part 1. *Comm. Pure Appl. Math.* 8:13–46.

Neyman, J. 1976. Tests of Statistical Hypotheses and Their Use in Studies of Natural Phenomena. *Comm. Stat. Theor. Methods* 8:737–751.

Pearson, E. S. 1947. The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 × 2 Table. *Biometrika* 34:139–167. Reprinted in Pearson (1966).

Pearson, E. S. 1950. On Questions Raised by the Combination of Tests Based on Discontinuous Distributions. *Biometrika* 37:383–398. Reprinted in Pearson (1966).

Pearson, E. S. 1955. Statistical Concepts in Their Relation to Reality. *J. Roy. Stat. Soc.*, ser. B, 17:204–207.

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

Pearson, E. S., and Neyman, J. 1930. On the Problem of Two Samples. *Bull. Acad. Pol. Sci.*, 73–96. Reprinted in Neyman and Pearson (1967).

Royall, R. M. 1997. *Statistical Evidence: A Likelihood Paradigm.* London: Chapman and Hall.

Royall, R. M. 2004. The Likelihood Paradigm for Statistical Evidence. Chapter 5 in Taper, M. L., and S. R. Lele, eds., *The Nature of Scientific Evidence: Empirical, Statistical, and Philosophical Considerations.* Chicago: University of Chicago Press.

Savage, L. J. 1972. *The Foundations of Statistics.* New York: Dover.

Spanos, A. 2000. Revisiting Data Dining: "Hunting" with or without a License. *J. Econ. Methodology* 7:231–264.

5 The Likelihood Paradigm for Statistical Evidence

Richard Royall

ABSTRACT

Statistical methods aim to answer a variety of questions about observations. A simple example occurs when a fairly reliable test for a condition or substance, C, has given a positive result. Three important types of questions are: (1) Should this observation lead me to believe that C is present? (2) Does this observation justify my acting as if C were present? (3) Is this observation evidence that C is present? We distinguish among these three questions in terms of the variables and principles that determine their answers. Then we use this framework to understand the scope and limitations of current methods for interpreting statistical data as evidence. Questions of the third type, concerning the evidential interpretation of statistical data, are central to many applications of statistics in science. We see that for answering them current statistical methods are seriously flawed. We find the source of the problems, and propose a solution based on the law of likelihood. This law suggests how the dominant statistical paradigm can be altered so as to generate appropriate methods for (i) objective representation and measurement of the evidence embodied in a specific set of observations, as well as (ii) measurement and control of the probabilities that a study will produce weak or misleading evidence.

INTRODUCTION

An important role of statistical analysis in science is interpreting observed data as evidence—showing "what the data say." Although the standard statistical methods (hypothesis testing, estimation, confidence intervals) are routinely used for this purpose, the theory behind those methods contains