Experimental practice, if viewed as more than a source for experimen­
tal narratives or illustration, could produce a decisive transformation
in the image of scientific inference by which we are now possessed.1
Anyone who looks upon the nitty-gritty details of most experimental
testing in science soon discovers that our neat and tidy models of
experimental inference have very limited applicability. The data are
inexact, noisy, and incomplete; extraneous factors are uncontrolled or
physically uncontrollable; it may be impossible to replicate and nearly
so to manipulate; and there may be huge knowledge gaps between the
scope of theories and any experiment we can actually perform. Theo­
ories, even theories plus auxiliaries, fail to tell us how to test them and
are silent about what to expect in the experiment before us. Yet despite
these limitations, we do often manage to obtain reliable knowledge
from data, and a question of interest to philosophers of science is
“How?” How do we learn about the world in the face of limited
information, uncertainty, and error?

An Adequate Account of Learning in the Face of Limits and Error

Is it possible to have a general account of scientific inference or testing
that shows how we learn from experiment despite these limitations?
One way that philosophers have attempted to answer this question
affirmatively is to erect accounts of scientific inference or testing in
which the limitations and uncertainties would be accommodated by appealing to probabilistic or statistical ideas. Leading attempts take the form of rules or logics relating evidence (or evidence statements) and hypotheses by measures of confirmation, support, or probability. Granting that hypothetico-deductive models are too limited to account for scientific inference, many propose that science can have a logic based on probability theory. We can call such accounts *logics of evidential relationship* or *E-R logics*. The leading example of such an E-R logic on the contemporary philosophical scene is the subjective Bayesian approach. Beginning with an accepted statement of evidence $e$, scientific agents are to assign prior probabilities, understood as measuring subjective degrees of belief, to an exhaustive set of hypotheses, which are then updated according to Bayes's formula from the probability calculus. What we have learned about a hypothesis $H$ from evidence $e$ is to be a function of the difference between the agent's posterior degree of belief in $H$ given $e$ and the prior degree of belief in $H$. Take Howson and Urbach: "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 which, from the point of view of the theory, are simply irrelevant" (Howson and Urbach 1989, 272; emphasis added). This is the Bayesian's way of doing something akin to deductive logic while taking into account the limitations of data in scientific inquiry.

The increased interest in studying actual experimental episodes in science, however, has led many philosophers to regard such logics of evidential relationship as failing to do justice to the actual limits and errors of inquiry. Scrutinizing evidence, in practice, is not a matter of logical or probabilistic relationships between statements, but turns on *empirical* information about how the data were generated and about the overall experimental testing context. It depends not on appraising full-blown theories or hypotheses against rivals but rather on *local experimental tests* to estimate backgrounds and to distinguish real effect from artifact and signal from noise. Philosophers, by and large, have either despaired of coming up with a systematic account of evidence and testing or else pursued minor tinkering with the existing logics of evidential relationship. The former group has concluded that the complexities and context-dependencies of actual experimental practice seem recalcitrant to the kind of uniform treatment dreamt of...
by philosophers; the latter seems to have accepted the false dilemma of "Bayes or Bust."2

I urge a move away from both these paths and propose a third way. The "data" from experimental practice should serve, not just as anomalies for philosophical models of evidential support, but as evidence pointing to a substantially different kind of account of experimental testing. What the evidence from practice shows is that where data are inexact, noisy, and incomplete, there is often disagreement and controversy as to whether they provide evidence for (or against) a claim or hypothesis of interest. Thus an adequate account must not begin with given statements of evidence but should provide methods for determining if we even have evidence for a hypothesis to begin with. An adequate account should also be able to motivate the ways in which scientists actually struggle with and resolve disagreements about evidence in practice: and in these struggles it is clear that scientists are in need, not of a way to quantify their beliefs, but of a way to check if they are being misled by beliefs and biases in interpreting their own data and that of other researchers. Yet to perform this check requires something else that an E-R logic will not give them—it requires some way to assess reliability. In other words, even when the E-R logic has done its work—say high support for \( H \) is found—we still need to know something more: how frequently would your measure of support give this much credit to hypothesis \( H \) even it would be a mistake (to regard the data as good evidence for \( H \))? If it would often be a mistake, then the test has not passed a reliable test (the test lacks severity). The probability that a procedure will be wrong over a series of applications is called an error frequency or error probability. A philosophy of scientific inference based on error frequencies may be called an error-statistical account.

Nevertheless, simply pointing to the widespread use of error-statistical methods in scientific practice (e.g., Fisherian and Neyman-Pearson tests) is not yet to motivate my claim that they contain essential features for an adequate philosophy of evidence. Indeed, there has been a good deal of controversy as to the relevance of a test's long-run frequencies of error in interpreting evidence—leading many philosophers to prefer Bayesian and other E-R logics (see, e.g., Mayo 1996 and Mayo and Kruse, forthcoming). My goal in this chapter is both to elucidate key aspects of the error-statistical account that I favor and at the same time to offer a rationale for making error probabilities central.
to the interpretation of data. Focusing on just one key type of statistical tool (based on “null hypothesis” testing), I will show how error-statistical reasoning effectively deals with limited information in very different contexts.

**Some Examples**

Consider these examples:

- Is the observed correlation between a given (ApoE4) gene and Alzheimer’s disease (AD) evidence for a genetic theory of AD (as against the leading neurological theories)?
- Is the failure of a spectrometer to detect muons in a proportion of events evidence of the existence of neutral currents?
- Are observed numbers of extinctions in given stages of the fossil record evidence of mass extinction events (as postulated by a given theory in paleontology)?

In each case the evidence accords with or “fits” a given hypothesis against a rival. One might express this by means of one of the E-R measures of fit, say, by asserting that a hypothesis $H$ makes $e$ more likely than do (extant) rivals. Although the E-R logic has done its job, the work actually required to answer the question “Is $e$ evidence for $H$?” is not thereby ended but has only just begun. Something beyond the measures of “fit” offered by E-R logics is required, or so I claim, in order to tackle this question. Two pieces of data that would equally well support 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, scientists, it seems, seek to scrutinize if the overall experiment from which the data arose was a reliable probe of the ways we could err in taking $e$ as evidence for (or against) hypothesis $H$—that is, the ways we can mistake $e$ as evidence for $H$. 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.

**Modeling Mistakes**

This does not mean we have to give up saying anything systematic and general, as many philosophers nowadays fear. The hodgepodge of
methods gives way to rather neat statistical strategies, and a handful of similar models may be used to probe a cluster of mistakes across a wide variety of domains. Although limited information may lead to uncertainty as to whether we have any kind of evidence for a hypothesis, we may know a good deal about how the type of evidence can be mistaken as evidence for $H$. On this knowledge, we base models of mistaken construals of evidence.

The mistakes arise because although the data may accord with a hypothesis, we cannot be sure they are not actually the result of “background” or “noise,” of artifacts for which we have not controlled, or of faulty experimental and theoretical assumptions. Rather than be stymied by our limited control, we may instead learn enough about background factors to “subtract them out” in comparing hypotheses to data or to estimate the likely upper bound of their influence. In coping with potentially faulty assumptions whose influences we cannot estimate, we may deliberately run tests with varied assumptions and check for converging results.

So what we need are models and methods for estimating what it would be like were it a mistake to regard $e$ as evidence for $H$, and strategies for discerning whether the actual situation is one of these mistaken ones. The history of mistakes made in a type of inquiry gives rise to a list of mistakes that researchers either work to avoid (before-trial planning) or check if committed (after-trial checking). For example, when inferring the cause of an observed correlation, such a repertoire of errors might include a set of questions: Is the correlation spurious? Is it due to an extraneous factor? Are we confusing cause and effect? Corresponding to such a repertoire of errors is a “reservoir of models.” I call them models of error.

### Null Hypotheses

A standard source of models of error may be found in what are commonly called null hypotheses or null models in statistics. These null, or error, hypotheses let us model, in effect, the situation of being fooled into thinking $e$ is evidence for $H$. Hence we have a warrant for taking $e$ as evidence for $H$ only to the extent that we can reject the null hypothesis. The term null comes from the fact that the situation of being fooled may often be expressed as there is really “no effect,” whereas an
alternative hypothesis $H$ asserts that some real effect or phenomenon exists. (We can more generally call them error hypotheses, but I will usually keep to the familiar term here.)

Take the classic null hypothesis that says a given pattern (in $e$) is or may be regarded as nongenuine or merely "due to chance." By providing us with a contrast against which to compare the observed pattern, the null model $6$ lets us see if the data can easily be accounted for by mere chance. These null or error models provide ways to show just how easily (i.e., frequently) results that may appear to show the absence of the error can be produced even when the error is present. Tests can be designed so that with high probability they would yield a result deemed "reasonably typical of a process in which the error is committed," if in fact it is committed—but not otherwise. Thus if a result is one that the test deems practically incapable of arising under the assumption of error, it is a good indication that the error is absent. Many different kinds of problems in learning from limited data may be tackled by comparisons with a null or error model, and a considerable amount of effort goes into constructing useful null models. Models of error are not limited to statistical models. However, because the nonstatistical variants may often be seen as attempts to approximate statistical models, understanding the statistical case is a useful springboard for extracting a general kind of experimental argument.

**Arguing from Error**

The argument I have in mind follows an informal pattern of reasoning that I call an argument from error or learning from error. The overarching structure of the argument is guided by the following thesis, which I will state in two equivalent ways:

(a) It is learned that an error is absent when (and only to the extent that) a procedure of inquiry signals the absence of the error, despite the procedure having a very low probability of doing this if in fact the error is present.

That a test signals the absence of the error means that the test produces a result that is classified by the test as "no error." The job of a statistical test is to indicate which of the possible test results to classify as indicating (or signaling) the absence and which results to classify as indicating the presence of the error of interest, and to do so in such a way as to substantiate the above kind of argument from error. Since the set of results classified as indicating "no error" is the complement
of the set indicating the presence of the error, the above argument from error can be written equivalently as follows:

\[(b) \quad \text{It is learned that an error is absent when (and only to the extent that) a procedure of inquiry signals the absence of the error, despite the procedure having a very high probability of signaling the presence of the error, if the error is present.}\]

Although form (a) of the argument from error may seem easier to parse, the equivalent form (b) brings out why such a procedure of inquiry may be called a reliable or highly severe error probe. The value of this “high probability” would measure the severity with which the test result indicates the error is absent. According to this thesis, we can argue that an error is absent if it fails to be detected by a highly severe error probe. A corresponding argument may be given to learn that an error is present, and indeed, one of the best-known “null experiments” in science, that of Michelson and Morley, involved identifying the presence of an error by failing to reject a null hypothesis.

The pattern of arguing from error underlies many different experimental arguments. The kind of example on which I focus can be described in terms of a test of a hypothesis \( H \), where hypothesis \( H \) asserts that a given error is absent—or is less than a given amount. Correspondingly, a null or error hypothesis, often written as \( H_0 \), may be used to express the presence of the error (i.e., not-\( H \)). Experimental results are good evidence for \( H \) to the extent that \( H \) passes a severe test with these results—that is, a test with a high probability of failing \( H \) (and detecting the presence of the error) just in case \( H \) is false (and the error is present, i.e., \( H_0 \) is true).

To elaborate by means of a familiar example, suppose we are testing whether a coin is biased for “heads” on the basis of observing \( e \), the proportion of heads in one hundred tosses (appropriately made). The null hypothesis \( H_0 \) is:

\[H_0 \quad \text{It is an error to take } e \text{ as due to genuine bias (i.e., any observed discrepancy from 50 percent heads is merely “due to chance”).}\]

while \( H \) asserts:

\[H \quad \text{It is not an error to take } e \text{ as due to genuine bias.}\]

If the test classifies results between 30 and 70 percent heads as still consistent with the presence of the error, that is, as still not so far from 0.5 to indicate bias, then the test will correctly signal the presence of
the "error" more than 99.9 percent of the time. It is a severe error probe (where the error consists of taking the coin as biased for heads when it is actually fair). Thus, an outcome \( e \) such as 71 percent heads would be classified as signaling \( H \), and \( H \) would thereby be said to pass a severe test with \( e \).

A question that immediately arises is whether such severe tests may be constructed for more general hypotheses in science. In particular, an important challenge in setting out this notion of severity is to show how it avoids the obstacles that vitiate other, broadly analogous, notions of severity (e.g., that of Popper). The most serious obstacle is this: how can one satisfy the severity requirement—that there be a high probability of detecting \( H \)'s errors if \( H \) is false (i.e., \( H_0 \) is true)—when there are always alternatives to \( H \) that have not even been thought of? What enables my account of severity to answer this objection is that it is a piecemeal account of testing. Where we cannot test everything at once, we may be able to test piecemeal—and it is precisely the role of null models, together with a methodology of testing, to enable such piecemeal testing to proceed.

**Models of Inquiry**

For each experimental inquiry we can delineate three types of models (figure 19.1): models of primary scientific hypotheses, models of data, and models of experiment that link the others by means of test procedures.

A substantive scientific inquiry is to be broken down into one or more local primary hypotheses, and the experimental models serve as the key linkage models connecting the primary model to the data models. To determine if data provide strong evidence for a hypothesis requires considering how reliably a test is able to detect a given type of error, that is, the test's error probabilities or error characteristics. The need for one or more null models or hypotheses arises to give us some estimate of these error probabilities, that is, to assess how good the test is at revealing if the situation is one of the erroneous ones. The null

---

**Figure 19.1** Three models of inquiry.
model is defined within the experimental model for the test, which is very often statistical.

The thrust of experimental design is deliberately to create contexts that enable questions to be asked one at a time. Within an experimental testing model, a null hypothesis asserts that “H is false,” or the situation is as if H is false. “H is false” (H₀ is true) refers to a specific error that the hypothesis H is denying. If H states that a given effect is systematic—of the sort brought about more often than by chance—then not-H states that it is due to chance; if H states that a parameter is greater than some value c, not-H states that it is less than c; if H states that factor F is responsible for at least p percent of an effect, not-H states that it is responsible for less than p percent, and so on. In each case, the not-H may be expressed as a null hypothesis H₀ that asserts it would be an error to take the data as evidence of H. More than that, H₀ may be used to model what it would be like if it were an error to take the data as evidence of H. How specific the error hypothesis is depends upon what is required to ensure a good chance of learning something of interest.

Looking at the problem in terms of the logical or probabilistic relationships between given evidence and hypotheses overlooks all of the active intervention that provides the basis for arguing that if a specific error is committed, it is almost certain to show up in one of the results of a given probe or series of tests—each with deliberately varied assumptions. By such active intervention one can substantiate the claim that we should expect (with high probability) hypothesis H to fail a given test, if H is false. Equivalently, there is a high probability of accepting the error explanation asserted by the null hypothesis, if (and only if) in fact the error is committed. This “active intervention” need not require any literal manipulation—far from it. The key service performed by the null hypothesis is to let us model what would be expected assuming the error when this necessitates a hypothetical or simulated set of possibilities. We will see this in our examples. So long as the effects of different factors can be sufficiently distinguished or subtracted out, then the inferences are not threatened by our limited ability to control for them.

**Standard (Error) Statistics**

The key value of standard error-statistical methods, as I see them, is that they provide very effective tools for modeling what would be expected (statistically) given that we are erring about some aspect of
the underlying cause or process of interest. They do so, first, by providing standard or canonical hypotheses to model the presence of an error (e.g., by means of standard null hypotheses) and, second, by providing tests that make it very hard, or extremely improbable, for us to rule out the error incorrectly. In referring to error statistics I include the familiar techniques of statistical analysis we read about every day in polls and studies, such as statistical significance tests and confidence interval estimates, although I adapt them in ways that go beyond what is strictly found in statistics texts.9 (Here I limit myself to discussing the former.) The name error statistics or error probability statistics comes from the fact that what fundamentally distinguishes this approach from others is that to determine what inferences are warranted by evidence requires considering the error probabilities of the overall testing procedure. (For an in-depth discussion of how this distinguishes error statistics from the Bayesian approach, see Mayo and Kruse, forthcoming.) In other words, it requires, in addition to a measure of how well e fits H, a measure of the probability of achieving so good a fit, even if H is false.

Anyone who opens an issue of Nature or Science sees significance tests and interval estimates used to scrutinize evidence in widely different fields. Yes, there are controversies surrounding the interpretation and justification of null hypothesis significance tests, but it seems to me that it is precisely the job of the philosopher of science to help resolve, or at least help clarify, these controversies. Certainly it has not helped matters for philosophers of science to dismiss the widespread use of standard statistical tests as the result of confusion and/or brainwashing (as Howson 1997 and some other Bayesian critics have alleged10). Putting aside the textbook rationale for significance tests (in terms of low long-run error rates) and their well-worn criticisms, I propose that philosophers undertake a serious reexamination of these tests and consider whether—and if so how—they may be serving to provide, if not a uniform evidential-relation logic, a valuable set of tools for coping with limitations, uncertainties, and errors in learning from data.

To illustrate the kind of examination I have in mind, let us consider aspects of the three examples delineated earlier.

Alzheimer’s Disease

Suppose we have recorded that a high proportion of people suffering from a disease, say AD, are found to have some factor, such as the
presence of a special gene (E4). This is our data \(e\). Is this evidence for a genetic theory of Alzheimer's disease?

The Bayesian rule asks us to compute \(P(\text{false} | e, \text{given the genetic theory is false})\), but our information is much too limited for us to proceed with such a computation. Even restricting “the genetic theory is false” to the existing alternative theory (at the time of this evidential appraisal), the computation requires us to know the probability of \(e\) given the nongenetic, neurological theory of AD. But that theory says nothing about genes, so how can we even begin? The error statistician, by contrast, proceeds by considering, not alternative Alzheimer's theories, but alternatives to the claim that \(e\) is good evidence for the given (genetic) Alzheimer's theory.

In other words, the error statistician begins by recognizing that although a genetic hypothesis may accord with the observed proportion \(e\), in fact \(e\) might be very poor evidence for \(H\) because the inquiry that brought forth \(e\) has not even begun to rule out several ways that \(H\) can be in error. These errors can be stated and probed quite apart from a specific alternative theory of AD. The very limitation of the available evidence at this stage alerts the error statistician to the absence of a comparative group: that is, the observed proportion \(e\) gives no comparison of the proportion (with this genetic factor) among those who are not afflicted with AD. As one set of researchers put it: “Since the ApoE4 allele is . . . common . . . in the general population . . . one would expect that a substantial proportion of cases of the coincidence of AD and inheritance of the ApoE4 allele is due to chance and not inherited predisposition to the disease” (Tanzi et al. 1996).

They could, however, appeal to a model of “chance”: a model of what it would be like to err in taking the observed proportion as evidence of a genuine connection (between the allele and AD). More specifically, they could represent the proportion \(e\) that would be expected to have this gene if they were observing non-AD sufferers (in the given experiment). But they could not strip away their disease and see if they still have the gene. What they could do (and did) was to obtain an appropriate comparison group all of whom lack AD—a control group—and observe the proportion of the suspect gene among them. All of this is very familiar, and I begin with such a familiar point to facilitate our understanding of some less familiar ideas.

We can employ a standard “null” hypothesis \(H_0\) to assert: there is no genuine correlation between the disease and the presence of the special gene (E4). Equivalently, \(H_0\) says it is an error to suppose a
genuine correlation is responsible for any observed correlation. It is not enough to report a difference between the two proportions. The null model must enable us to say, approximately, how often there would occur a difference in proportions (with the gene) as large as the difference observed, if in fact $H_0$ were the true state of affairs (i.e., if taking the result as evidence of a real correlation would be a mistake). That is, the null model must give us an error probability; and in particular, it must give us what is called the statistical significance level of the observed difference in proportions.

However, in order for models of error to give us error probabilities, there typically must be a deliberate introduction of statistical considerations into the data generation (e.g., by randomization, by approximations using matched controls) or by means of the data modeling (e.g., as with simulations). Only then can we carry out the comparison with the null model, by calculating the statistical significance of the data.

**The Significance Question**

Picturing our scheme of models, we can locate the hypothesis $H$, about the correlation, in an experimental model of an inquiry. To probe $H$, in our example, the researchers developed a design that would let them use the standard statistical null hypothesis $H_0$, which approximates to the familiar normal distribution of error. This allowed the researchers to direct a question to the given observed difference, which we may abbreviate as $d$. The question they posed of $d$ may be called a significance question. It is this:

*How frequently would we obtain a result that accords with $H$ as well as $d$ does (or better) if in fact $H$ is false and $H_0$ is true (i.e., the error is committed)?*

What is being asked, in statistical terminology, is: what is the statistical significance level of the observed difference $d$? The null model is what let the researchers answer this question. The answer (i.e., the statistical significance level) is of interest for the simple reason that they want to be assured—before they take the observed fit (between $d$ and hypothesis $H$) as evidence of a real effect—that $d$ would not be expected to occur fairly frequently (in samples of this size) if null hypothesis $H_0$ were true. So if the significance level says it is fairly frequent—that is, if the significance level is not very small—the researchers do not take the data as ruling out the null or error explanation. But in this particular case study, the significance levels observed were very small, that is, they were statistically significant.
To illustrate, suppose a statistically significant correlation is observed between those with AD and the suspect gene: perhaps the significance level (or p-value) is .01. This report might be taken to reject the null hypothesis \( H_0 \) and imply there is a genuine correlation between the two. Notice that in rejecting \( H_0 \) we are rejecting or denying the error asserted in \( H_0 \). But the significance level of .01, with which we reject this error, is not an assignment of probability to the null hypothesis—we cannot say \( H_0 \) has probability .01. Rather, .01 is the probability that such a test procedure would reject the null hypothesis erroneously, thus the term error probability. It asserts, in particular, that were it an error to infer the genuine correlation (i.e., were \( H_0 \) true), so large an observed correlation (as we have in \( d \)) would occur only 1 percent of the time.

Hence, following the argument from error, the low significance level, .01, lets us pass the hypothesis \( H \), that the correlation is genuine. Evidence from any one test might at most be taken as evidence that the correlation is genuine. But after several such failed null hypotheses, hypothesis \( H \) passes a severe test because, were \( H \) false and each null hypothesis \( H_0 \) true, we would very probably (probability >.99) have obtained results that accord less well with \( H \) than the ones we got. Note that it is the entire procedure of the various subexperiments that may properly be said to have the probative power—the high probability of detecting a spurious correlation by not yielding such consistently statistically significant results. In this way, the significance test informed the researcher (Alan Roses at Duke University) whether he would be wasting his time trying to find the cause of something that might be regarded as accidental.

I admit that my discussion here remains sketchy and at a rudimentary level, but this should suffice for my present goal, which is to identify key features of how I think error-statistical reasoning enters to cope with limitations and errors in a cluster of cases. These inferences are, by and large, directed to models of experiment intermediate between the data and substantive scientific hypotheses or theories. Nevertheless, inferences about such experimental models are often taken as input into models of primary hypotheses of interest, thereby providing indirect evidence for hypotheses in the primary scientific model. One can of course attach many additional models to the scheme in this same fashion.

Information from one experimental model is often used to supply a piece of information that can be used as a clue for further investiga-
tions. For instance, having severely passed the correlation hypothesis between the gene E4 and AD, researchers had clues about a possible genetic explanation, but they had not yet reliably demonstrated a phenomenon that rival causal hypotheses had to confront. Not only did the genetic hypothesis of AD go against the neurological hypotheses generally accepted at the time (which regarded AD as caused by a buildup of plaque in the brain), the researchers had simply not yet ruled out numerous ways in which they would have been mistaken to infer a causal connection from such a correlation—however real that correlation was found to be. Specifying which errors are ruled out at a given stage also reveals which errors have not yet been considered; and that suggests what would be needed to go further in investigating the phenomenon of interest. In this case, it is evident that one must next go on to confront causal errors.

The first causal error they detected was that it was not ApoE4 itself that caused AD, but the lack of better versions of this gene (E3 or E2) that was responsible for the early death of nerve cells, which then produced the amyloid tangles. The tangles, it seemed, were an effect and not a cause of AD. This indicated that the rival amyloid plaque hypotheses might have the causal story, and even the causal order, wrong. 13

Using Data to Probe Hypotheses that Say Nothing about Them

As I remarked earlier, the neurological (plaque) theories said nothing whatsoever about genes (or the associated protein ApoE) involved in the causal process. They neither predicted nor counterpredicted the genetic evidence. But this evidence turned out to be probative for understanding what happens in the brains of Alzheimer’s patients, and thus for identifying (unexpected) anomalies in existing neurological hypotheses. (It was said that the link had never even been hinted at by any of the world’s foremost Alzheimer’s scientists.)

This recognition is important for seeing how the current account circumvents a familiar criticism of accounts of testing—one that Larry Laudan (forthcoming) has reemphasized. Theories of testing, Laudan alleges, cannot account for the fact that in practice theories are appraised by means of phenomena that they do not explicitly address. Such an appraisal, Laudan thinks, is outside the scope of testing, and he concludes that it requires some nonepistemic analysis, thereby showing the in-principle limits of epistemology.
I allow that Laudan’s criticism hits its mark with respect to Bayesian theories, which are his principal target, because, as noted earlier, it is not clear how one can assign the needed probability of the data given a rival hypothesis that says nothing about the data domain observed. Nevertheless, we have just illustrated how error-statistical testing avoids his charge. One way to express the current state of knowledge regarding a phenomenon or domain, I propose, is in terms of the errors that have and have not been ruled out by severe tests. Although certain aspects of the amyloid plaque theories had passed severe tests (for example, there is evidence of some genuine connection between these tangles and AD), none of the hypotheses about how amyloid causally contributes to AD had yet passed severe tests. This is why Roses was led to question the generally held causal story in the first place. So by passing even some aspects of his causal account severely (even lacking anything like a complete theory of AD), Roses was able to obtain evidence that the amyloid account gets the causal story wrong; and he was able to do this despite the fact that plaque theories said nothing about genes.

Again, I admit to being skimpy with the details of a case with an ever-changing evidential base, but my aim is just to show the power of the methods and logic of this account to circumvent obstacles due to common information gaps. It seems to me that a strikingly similar pattern of arguing from error occurs in widely diverse fields, often using the same models of error. Let us now consider an example from twentieth-century physics.

Distinguishing Effects from Artifacts: Galison and Neutral Currents

Although by the end of the 1960s, Peter Galison (1987) tells us, the “collective wisdom” was that there were no neutral currents (164, 174), soon after (from 1971 to 1974) “photographs . . . that at first appeared to be mere curiosities came to be seen as powerful evidence for” their existence (135). How did experimentalists themselves come to regard this data as evidence that neutral currents existed? “What persuaded them that they were looking at a real effect and not at an artifact of the machine or the environment?” (136).

To give the bare bones of the analysis, experimental outcomes here are modeled using a (dichotomous) statistic somewhat like that in the AD case. There each subject either had the E4 gene or did not. Here...
each event is described as muonless or muonful. In the AD case, the recorded result was the difference in proportions (with the suspect gene); here the recorded result is the ratio $R$ of the number of muonless and muonful events. (Neutral currents are described as those neutrino events without muons.) The main point is that the more muonless events recorded, the more the result favors the existence of neutral currents. The worry is that recorded muonless events are due not to neutral currents but to inadequacies of the detection apparatus.

Experiments were conducted by a consortium of researchers from Harvard, Wisconsin, Pennsylvania, and Fermilab, the HWPF group. They recorded 54 muonless events and 56 muonful events, giving a ratio of 54/56. Rubbia, a researcher from Harvard, emphasized that “the important question . . . is whether neutral currents exist or not. . . . The evidence we have is a 6-standard-deviation effect” (220). The “important question” revolved around the question of the statistical significance of the effect.

**The Significance Question**

The **significance question** in this case is: What is the probability that the HWPF group would get as many as (or more than) 54 muonless events, given there are no neutral currents? We could answer this if we had a null model to tell us how often, in a series of experiments such as the one performed by the HWPF group, we would expect the occurrence of as many muonless events as were observed, given that there are no neutral currents. This is precisely what the researchers sought to determine.

It may be objected that there is only this one experimental result, not a series of experiments. True, the frequency of outcomes in a series of experiments supplied by the null model is a kind of hypothetical construct—just as in the case of AD. I am trying to bring out why it is perceived as so useful to introduce this hypothetical construct into the data analysis.

The answer, as I see it, is that such a null model provides an effective way to model the distribution of outcomes assuming it was a mistake to regard the data as evidence of the existence of neutral currents. It thereby provides an effective way to distinguish real effect from artifacts (correct from incorrect interpretations of the given result). Were the experiment so well controlled that the only reason for failing to detect a muon is that the event is a genuinely muonless one, then
artifacts would not be a problem and this statistical construct would not be needed. But from the start a good deal of attention focused on the backgrounds that might fake neutral currents. A major problem was escaping muons. “From the beginning of the HWPF neutral current search, the principal worry was that a muon could escape detection in the muon spectrometer by exiting at a wide angle. The event would therefore look like a neutral-current event in which no muon was ever produced” (217).

The problem, then, is to rule out a certain error: construing as a genuine muonless event one in which the muon simply never made it to the spectrometer and thus went undetected.

If we let hypothesis $H$ be

$$H \quad \text{Neutral currents are responsible for (at least some of) the results.}$$

then, within this piece of data analysis, the falsity of $H$ is the artifact explanation:

$H$ is false (the artifact explanation): Recorded muonless events are due, not to neutral currents, but to wide-angle muons escaping detection.

The null hypothesis asserts, in effect, that the observed ratio arose from a universe where $H$ is false. Using the familiar abbreviation, $H_0$, we have

$$H_0 \quad \text{It would be an error to regard the observed ratio as evidence of neutral currents.}$$

Our significance question becomes: What is the probability of a ratio (of muonless to muonful events) as great as 54/56, given that $H$ is false (that is, given $H_0$)? The answer is the significance level of the result.

But how do you get the significance probability or even approximate it? It requires estimating what would be usual or typical (statistically speaking) due to the background alone. The HWPF group, for example, created a computer simulation called a Monte Carlo program in order to model statistically how muons could escape detection by the spectrometer by exiting at a wide angle. “By comparing the number of muons expected not to reach the muon spectrometer with the number of measured muonless events, they could determine if there was a statistically significant excess of neutral candidates” (Galison, 217).

Note that probability arises in this part of the analysis not because the hypothesis about neutral currents is a statistical one, much less because it quantifies credibility in $H$ or in $H_0$. Probabilistic consider-
ations are deliberately *introduced* into the data analysis because they offer a way to model the expected effect of the artifact (escaping muons). In the AD example, we saw how statistical considerations were introduced by specially selected control groups. Here they are introduced by considering a statistical simulation of the background by a computer. Statistical considerations—we might call them "manipulations on paper" (or on computer)—afford a way to subtract out background factors that cannot literally be controlled for.

**The Data**

The data used in the HWPF paper are as follows (Galison, 220):

- Visible muon events: 56
- No visible muon events: 54
- Calculated muonless events: 24
- Excess: 30
- Statistically significant deviation: 5.1

The first two entries just record the HWPF result. The third entry refers to the number calculated or expected to occur because of escaping muons, as derived from the Monte Carlo simulation.

The simulation can be seen as allowing us to construct a null model: a model of the relevant features of what it would be like (statistically) if the researchers were actually experimenting on a process in which the artifact explanation $H_0$ is true. It tells us, in particular, that given outcomes (observed ratios) would occur with certain probabilities. (Most experiments would yield ratios close to the average (24/56); the vast majority would be within two standard deviations of it.) The null hypothesis, in effect, asserts that the observed result—54 out of 56—is *not* out of the ordinary even if *all* events are due to escaping muons. The difference between the ratio observed and the ratio expected (due to the artifact) is $54/56 - 24/56 = 0.536$. The null model tells us how improbable such a difference is even if the experiment were being performed on a process in which the artifact explanation is true (i.e., in which recorded muonless events were all due to escaping muons). In this way it provides the statistical significance level that was sought.

Putting an observed difference between recorded and expected ratios in standard deviation units allows one to use a chart to read off the corresponding error probability. Any difference exceeding two or more standard deviation units is one that is improbably large (occu-
ring less than 2 percent of the time). Approximating the standard deviation of the observed ratio showed the observed difference to be 5.1 standard deviations! This is so improbable as to be off the charts; so, clearly, by significance-test reasoning, the observed difference indicates that it is practically impossible for so many muonless events to have been recorded, were they all due to the artifact of wide-angle muons. The procedure is a highly reliable artifact probe.

Only after varying the analysis in many ways was the hypothesis of a real effect regarded as having passed a severe test. For example, Galison explains, the researchers deliberately used three distinct methods to calculate the ratio $R$, and “since each of the three methods used the Monte Carlo program in very different ways, the stability of the data suggested that there was no gross error in the subtraction method” (219). In high-energy physics, as on the laboratory bench, Galison tells us, “the underlying assumption is the same: under sufficient variation any artifact ought to reveal itself by causing a discrepancy between the different ‘subexperiments’” (219). And even this multivaried analysis was just one small part of a series of experimental arguments for the existence of neutral currents that took years to build up. My point is that each involved this kind of statistical reasoning to distinguish real effects or signals from artifacts, to estimate the maximum effect of different backgrounds, and to rule out key errors piece-meal. They were put together to form the experimental arguments that showed that the experiment could end (to allude to the title of Galison’s book).

**Learning despite Limits in Paleobiology**

In the previous two examples the statistical models needed to answer the “significance question” in analyzing data were arrived at either by techniques of data generation or by special computer simulations. The need for, and the ability to apply, each technique was a function of the particular limits in the available evidence with which researchers were confronted. In the current example the limitations are even more severe. Here knowledge of the statistical distribution of outcomes (given that the error of interest is present)—the knowledge that the statistical null model is to provide—cannot be obtained in either of the ways illustrated earlier. Answering the significance question may still be possible, however, thanks to some newer (computer-driven) analytical
techniques. With these newer techniques, there is an attempt actually to construct the statistical distribution of possible results, given that the error is present (i.e., given that one is sampling from a population in which the error is committed). Doing this often takes "brute force," and indeed a method of growing popularity is sometimes called the "brute force" method.

Paleontology offers good examples to illustrate. Here we have a science limited by the inability both to replicate and to manipulate variables in any literal way: "We have only one [fossil record for a given period]; we cannot replay [it] twenty times to see how much variation could have been produced by the system. . . . However, these difficulties do not lead us to abandon the scientific approach to the fossil record." What prevents them from abandoning the scientific approach, they explain, is their ability to use to a "toolkit" of statistical techniques in order "to evaluate the statistical significance of historical patterns without replicate samples or repeated experiments" (Singer and Gilinsky 1991, 2).

Resampling Statistics

One technique that is increasingly used is resampling statistics (Efron 1979). Thanks to the use of computers, this new method enables the generation, on the basis of a given finite sample, of a population of data sets that could have resulted in repeated sampling from the given set. Rather than employing a parametric statistical model (e.g., a normal distribution), the idea is to use the data themselves to generate the hypothetical (statistical) distribution needed for the null model, and in this way obtain the corresponding assessment of statistical significance.

Resampling is one of the more imaginative ways by which researchers appeal to error-statistical ideas to circumvent limitations of knowledge: if we cannot (literally) manipulate or replicate, we may still be able to simulate, and thereby learn what it would be like to do those things. This same theme underlies all of the error-statistical techniques and arguments that I am trying to articulate by means of examples.

Let us consider a specific application of resampling. To obtain evidence of a mass extinction—understood as an unusually steep increase in extinction rates during a given period—scientists appeal to a simulation model of the expected or "background" extinction rates. The
simulation is based on repeatedly resampling values from part of the fossil record. This simulation serves as the null (or error) hypothesis to which data can be compared.

In order to illustrate their method, the researchers describe a very simple, artificial case in which three taxonomic orders are observed and the number of familial extinctions recorded at different stratigraphic stages (figure 19.2). In this illustrative example, the highest number of extinctions was found to occur at stage 20 (i.e., 13 extinctions), and the researchers need to ask: is this evidence of a mass extinction stage? “To decide whether stage 20 should be characterized as a mass extinction stage, we need to ask whether this number of familial extinctions, 13, is [statistically] significantly larger than the number . . . that would normally be expected to occur during stage 20, given the extinction histories of the three orders that were present” (Hubbard and Gilinsky 1992, 153).

To answer this question, they consider the null hypothesis that stage 20 was actually not unusual, and that the 13 extinctions occurred during that stage “by chance” (Hubbard and Gilinsky 1992, 153). That is to say, the null hypothesis, $H_0$, asserts:

$H_0$ It would be an error to take the 13 extinctions as good evidence of a mass extinction at stage 20.

Hypothesis $H_0$ asserts, in effect, that stage 20 could just as well have experienced the other extinction values observed during the history of each order. But now we need a probability assignment for each of the possible extinction values (i.e., each set of three numbers—one for each order), and we are assuming, recall, that we cannot obtain this by relying on one of the known statistical distributions.

To this end, the computer will generate the distribution for us. It chooses an extinction number randomly and independently from each of the three orders, and adds them to get the total number of extinctions that could have occurred during stage 20 (or during any other stage in which all three orders exist). This number becomes the first “bootstrapped” number for generating the distribution needed (see note 16). The computer then chooses a random extinction value again from each of the three orders, sums them, and gets the second bootstrapped value. The researchers performed 10,000 such bootstraps. We can graph (see figure 19.2) the frequencies for each sum, yielding the distribution that will serve as the null or error model. (It is a
Figure 19.2 Schematic depiction of the bootstrapping method as applied to the problem of mass extinctions. The “observed” total numbers of familial extinctions for hypothetical stages 11–27 are shown in the second row of numbers. Below this row are three rows of numbers depicting the “observed” numbers of familial extinctions for hypothetical orders A, B, and C, each of which contributed a portion of the total number of extinctions. Under the bootstrapping method, one extinction number is chosen randomly and with replacement from each of the orders, and these numbers are summed to create bootstrapped extinction numbers for the stage. The creation of two such numbers is shown. By repeating the process many times (they repeat the process 10,000 times), a distribution of bootstrapped extinction numbers (shown on the right side of the figure) is built up. This can be thought of as the distribution of possible extinction numbers for the stage, given the known extinction histories of the orders in existence during the stage. The observed number of extinctions for the stage under scrutiny (in this case 13 extinctions occurred during stage 20; see the text) is then compared to the distribution of bootstrapped numbers to decide whether the observed number of extinctions could readily occur by chance. The observed number of familial extinctions in stage 20, which was 13, is shown by an arrow on the graph of the distribution, and the substantial amount of area in the tail of the distribution bounded by the observed value indicates that the observed number of extinctions could indeed occur readily by chance. Therefore, no mass extinction is indicated. (From Signor and Gilinsky 1991. Reprinted by permission.)
The significance question in this case asks: is the observed extinction value (13) statistically unusual assuming the extinction values are distributed randomly in time? By finding the frequency with which bootstrapped extinction values equal or exceed 13, the computer answers this question; that is, it tells us the significance level of the observed outcome, 13.

Only if the observed frequency (13) occurred fewer than 100 times in the 10,000 simulations did the researchers consider it statistically significant. In this illustrated case, however, the significance level for 13 was about .1. Hence the researchers do not reject the null hypothesis. Rather, they regard the data as "demonstrating that the observed number of extinctions could occur sufficiently often by chance that stage 20 should not be regarded as a stage of mass extinction" (Hubbard and Gilinsky 1992, 155).

The researchers also applied this new analytic method to reanalyze some older data interpretations. They found that two of the five stages that previous researchers had considered indicative of mass extinctions were, according to their reanalysis, easy to explain by the hypothesis of chance error. Whether these past studies really had unearthed evidence of mass extinctions had been controversial, and the controversy was dealt with by means of this innovatively derived null model. As with our earlier examples, uncertainties in the null models led the researchers to perform the analysis using deliberately varied assumptions. In order for a stage to be regarded as showing evidence of a mass extinction, they required it to be found statistically significant using four different null models with 10,000 resamplings each. (They classified three stages as indicating mass extinctions.)

Granted, in this example and the others there is plenty of (subject-specific) background information that is called upon to implement error-statistical tests, but my point is to show that the reasoning in each case follows the standard pattern of arguing from error: we want to assess whether a set of observations is statistically distinguishable from what would be expected were it generated from a null process.
(here, from the "background" extinction numbers). If the data could be seen as readily (i.e., frequently) generated by this null process, then it would be an error to construe the data as evidence of the hypothesis of interest, $H$ (e.g., mass extinction). If the specific outcome is expected to occur fairly frequently under this null or error process, then it does not constitute evidence for $H$.

R. A. Fisher had said, "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (Fisher 1947, 14). Echoing Fisher, we might say: we have evidence of the phenomenon described in $H$ when we know how to conduct an experiment that will very rarely fail to reject the null hypothesis—that is, very rarely fail to reject the claim that we are wrong (to infer we have evidence for $H$). Although this seems like a lot of negatives—and I suppose it is—a little practice shows that the most mundane day-to-day examples of learning from error are actually instantiations of this reasoning.\(^{17}\)

### Two Important Points

Our examples illustrate two points that relate directly to central disputes in the philosophy of statistics. First, it is clear in these examples that the standard provided by a low significance level is of use not simply because we want to avoid too often interpreting data erroneously in the long run of experience. Rather, it is of use because of its role in informing us of the process that produced the data in front of us. We use it as a standard to weed out erroneous interpretations of data. If I am correct that the rationale for controlling a test's error probabilities is to have a standard for reliably interpreting evidence, a long-standing set of criticisms and confusions surrounding null hypothesis significance testing will have been resolved.

Second, note that in each case arguing from error turns on considering outcomes other than the ones actually observed. To use the observed data to learn about their underlying cause we need to think beyond what happened to occur in this case to consider what else might have occurred. By contrast, logics of evidential relationship (e.g., Bayesianism) discount this consideration of outcomes other than the one observed for the purpose of reasoning from the data: \(^{18}\) "The question of how often a given situation would arise is utterly irrelevant
to the question how we should reason when it does arise” (Jaynes 1976, 247). But for those who use error statistics, reasoning from the result that did arise is crucially dependent upon how often it would arise. Lacking such information prevents us from ascertaining which inferences can be reliably drawn from limited data. It follows that if the goal is reliability, error-statistical methods appear in a much more favorable light than those approaches that do not take account of a procedure’s error probabilities.

Conclusion

Where we cannot test directly a theory $T$—which contains various hypotheses and may err in different ways—it is often possible to test severely, instead, one or more hypotheses of error: hypotheses that indirectly model “what it would be like” were it a mistake to construe data $e$ as evidence for a specific hypothesis $H$. Whether it is by pointing to a statistical calculation, a pictorial display, or a computer simulation, the “what would it be like” question, as I see it, is answered by means of an experimental model that approximates the relative frequency with which certain results would occur in an actual or hypothetical series of experiments. This experimental construct serves as the null or error model. The actual data can then be compared with the data that would be expected assuming the error. By controlling, at some small value, the probability of rejecting the hypothesized error erroneously, one can achieve the goal of making it very difficult to take data $e$ as good evidence for $H$, when in fact that would be unwarranted. One can thereby learn about theory $T$, keeping track of the errors that have and have not yet been severely ruled out. Although limited evidence may prevent us from reliably discriminating between a primary scientific hypothesis and its substantive rivals, we may be able to discriminate between correct and erroneous interpretations of this data with respect to this hypothesis.

These standard or canonical models of error may be located somewhere between the substantive scientific hypotheses and the particulars of the experimental items and processes; they are part of the experimental models that link actual data to substantive primary claims. It is here that the philosopher of experiment might stand to develop a systematic account of evidence. Assessing whether a model
is good for the purposes of a null or error model may be quite distinct from the criteria used in testing the adequacy of some substantive model. By developing an account in which inductive inference in science is viewed as reliably distinguishing and ruling out errors, we may begin to resolve a cluster of problems of how to learn from limited evidence that confront philosophers as well as practitioners.

NOTES

1. This is, of course, an adaptation of the first sentence of Kuhn (1962).
3. The idea that statistics provides a “reservoir of models” comes from Erich Lehmann’s (1990) discussion of R. A. Fisher and Jerzy Neyman.
4. A fuller discussion of these models and the associated “error statistical” philosophy may be found in Mayo (1996).
5. Null models also occur outside statistics. But their uses in statistical testing provide apt analogues for the nonstatistical cases.
6. The null hypothesis is an assertion about a parameter in the corresponding null model.
7. One way to state the corresponding argument from error would be: “It is learned that an error is present when (and only to the extent that) a procedure of inquiry with a very high probability of rejecting the hypothesized error if the error is absent nevertheless does not reject the hypothesized error.”
8. Their null hypothesis might be expressed as: “It is an error to suppose that the effect of the earth’s motion through the ether is detectable.” In this case, there was a failure to reject the error hypothesis: the null hypothesis was accepted.
9. Error statistical tools include, but are not limited to, significance tests as developed by R. A. Fisher and testing and estimation methods as developed by Neyman and Pearson. The points on which I differ from the uses and interpretations advocated by these statisticians are discussed in Mayo (1996).
10. Howson concludes a recent article with this declaration: “Why it is taking the statistics community so long to recognize the essentially fallacious nature of NP [Neyman and Pearson] logic is difficult to say, but I am reasonably confident in predicting that it will not last much longer” (1997, 289). I respond to Howson in Mayo (1997).
11. One need not deny that an alternative theory might give one clues as to how one may err in taking e as evidence for the theory in question. But assessing the presence or absence of this error need not call for assessing P(e, given the alternative theory)—an assessment about which we may have no clue.
12. With even more limited information, they might instead appeal to what is known as Student’s distribution and the corresponding t-test.
13. Taking advantage of work in a different area—the biochemistry of the cholesterol-carrying protein called ApoE—it turns out that the gene for ApoE is
located in the very place where Roses had found the suspect E4 gene in families with AD. Having the E4 version of the gene (rather than the E3) caused nerve cells to die sooner, which in turn caused the plaques.


15. All references to Galison will be to Galison (1987).

16. Accordingly this procedure is also very commonly called “bootstrapping,” but it must not be confused with Glymour’s (1980) notion.

17. See, for example, Mayo (1996, Chapter 1).

18. Their doing so follows formally from the acceptance of the likelihood principle. This is discussed in detail in Mayo (1996, Chapter 10).

REFERENCES


Laudan, L. Forthcoming. “Epistemology, Realism, and Rational Theory Evaluation.”


