CHAPTER THREE

The New Experimentalism and the Bayesian Way

> [F]amiliarity with the actual use made of statistical methods in the experimental sciences shows that in the vast majority of cases the work is completed without any statement of mathematical probability being made about the hypothesis or hypotheses under consideration. The simple rejection of a hypothesis, at an assigned level of significance, is of this kind, and is often all that is needed, and all that is proper, for the consideration of a hypothesis in relation to the body of experimental data available.
>
> —R. A. Fisher, *Statistical Methods and Scientific Inference*, p. 40

> [T]he job of the average mathematical statistician is to learn from observational data with the help of mathematical tools.
>
> —E. S. Pearson, *The Selected Papers of E. S. Pearson*, p. 275

Since Lakatos, the response to Popper's problems in light of Kuhn has generally been to "go bigger." To get at theory appraisal, empirical testing, and scientific progress requires considering larger units—whole paradigms, research programs, and so on. Some type of holistic move is favored even among the many philosophers who consciously set out to reject or improve upon Kuhn. Whatever else can be said of the variety of holisms that have encroached upon the philosophical landscape, they stand in marked contrast to the logical empiricist approaches to testing that concerned setting out rules of linking bits of evidence (or evidence statements) with hypotheses in a far more localized fashion. For post-Kuhnian holists, observations are paradigm laden or theory laden and testing hypotheses does not occur apart from testing larger units.

The lesson I drew from Kuhn in the previous chapter supports a very different line of approach. The lesson I take for post-Kuhnian philosophy of science is that we need to go smaller, not bigger—to the local tests of "normal science." Having so reworked the activity of normal testing I will now drop the term altogether, except when a reminder of origins seems needed, and instead use our new terms *standard testing* and *error statistics.* But the aims of standard testing are still rather like those that Kuhn sets out for normal science.

I agree with critics of logical empiricism on the inadequacy of a theory of confirmation or testing as a uniform logic relating evidence to hypotheses, that is, the evidential-relationship view. But in contrast to the thrust of holistic models, I take these very problems to show that we need to look to the force of low-level methods of experiment and inference. The fact that theory testing depends on intermediate theories of data, instruments, and experiment, and that data are theory laden, inexact, and "noisy," only underscores the necessity for numerous local experiments, shrewdly interconnected.

The suggestion that aspects of experiment might offer an important though largely untapped resource for addressing key problems in philosophy of science is not new. It underlies the recent surge of interest in experiment by philosophers and historians of science such as Robert Ackermann, Nancy Cartwright, Allan Franklin, Peter Galison, Ronald Giere, and Ian Hacking. Although their agendas, methods, and conclusions differ, there is enough similarity among this new movement to group them together. Appropriating Ackermann's nifty term, I dub them the "New Experimentalists."

Those whom I place under this rubric share the core thesis that focusing on aspects of experiment holds the key to avoiding or solving a number of problems, problems thought to stem from the tendency to view science from theory-dominated stances. In exploring this thesis the New Experimentalists have opened up a new and promising avenue for grappling with key challenges currently facing philosophers of science. Their experimental narratives offer a rich source from which to extricate how reliable data are obtained and used to learn about experimental processes. Still, nothing like a systematic program has been laid out by which to accomplish this. The task requires getting at the structure of experimental activities and at the epistemological rationale for inferences based on such activities.

To my mind, the reason the New Experimentalists have come up short is that the aspects of experiment that have the most to offer in developing such tools are still largely untapped. These aspects cover the designing, modeling, and analyzing of experiments, activities that receive structure by means of statistical methods and arguments.

This is not to say that the experimental narratives do not include the use of statistical methods. In fact, their narratives are replete with applications of statistical techniques for arriving at data, for assessing the fit of data to a model, and for distinguishing real effects from artifacts (e.g., techniques of data analysis, significance tests, standard errors of estimates, and other methods from standard error statistics). What has not been done is to develop these tools into something like an adequate philosophy or epistemology of experiment. What are needed are forward-looking tools for arriving at reliable data and using such data to learn about experimental processes.

In rejecting old-style accounts of confirmation as the wrong way to go to relate data and hypothesis, the New Experimentalists seem to shy away from employing statistical ideas in setting out a general account of experimental inference. Ironically, where there is an attempt to employ formal statistical ideas to give an overarching structure to experiment, some New Experimentalists revert back to the theory-dominated philosophies of decision and inference, particularly Bayesian philosophies. The proper role for statistical methods in an adequate epistemology of experiment, however, is not the theory-dominated one of reconstructing episodes of theory confirmation or large-scale theory change. Rather their role is to provide forward-looking, ampliative rules for generating, analyzing, and learning from data in a reliable and intersubjective manner. When it comes to these roles, the Bayesian Way is the wrong way to go.

My task now is to substantiate all of these claims. In so doing I shall deliberately alternate between discussing the New Experimentalism and the Bayesian Way. The context of experiment, I believe, provides the needed backdrop against which to show up key distinctions in philosophy of statistics. The New Experimentalist offerings reveal (whether intended or not) the function and rationale of statistical tools from the perspective of actual experimental practice—the very understanding missing from theory-dominated perspectives on scientific inference. This understanding is the basis for both my critique of the Bayesian Way and my defense of standard error statistics.

One other thing about my strategy: In the Bayesian critique I shall bring to the fore some of the statisticians who have contributed to the debates in philosophy of statistics. Their work has received too little attention in recent discussions by Bayesian philosophers of science, which has encouraged the perception that whatever statisticians are doing and saying must be quite distinct from the role of statistics in philosophy of science. Why else would we hear so little in the way of a defense of standard (non-Bayesian) statistics? But in fact statisticians have responded, there is a rich history of their response, and much of what I need to say has been said by them.

I will begin with the New Experimentalism.

3.1 The New Experimentalism

Having focused for some time on theory to the near exclusion of experiment, many philosophers and historians of science have now turned their attention to experimentation, instrumentation, and laboratory practices.[1] Among a subset of this movement—the New Experimentalists—the hope is to steer a path between the old logical empiricism, where observations were deemed relatively unproblematic and given primacy in theory appraisal, and the more pessimistic post-Kuhnians, who see the failure of logical empiricist models of appraisal as leading to underdetermination and holistic theory change, if not to outright irrationality. I will begin by outlining what seem to me to be the three most important themes to emerge from the New Experimentalism.

1. Look to Experimental Practice to Restore to Observation Its Role as Objective Basis

Kuhn, as we saw in chapter 2, often betrays the presumption that where an algorithm is unavailable the matter becomes one of sociology. He proposes that problems that are usually put as questions about the nature of and warrant for theory appraisal be reasked as sociological questions that have "scarcely even been stated before" (Kuhn 1977, 212). Whether intended or not, this has invited sociological studies into the role that interests and negotiations play in constructing and interpreting data. Interviews with scientists have provided further grist for the mills of those who hold that evidence and argument provide little if any objective constraint.

A theme running through the work of the New Experimentalists is that to restore the role of empirical data as an objective constraint and adjudicator in science, we need to study the actual experimental processes and reasoning that are used to arrive at data. The old-style accounts of how observation provides an objective basis for appraisal via confirmation theory or inductive logic should be replaced by an account that reflects how experimental knowledge is actually arrived at and how it functions in science.

Peter Galison (1987) rightly objects that "it is unfair to look to experimental arguments for ironclad implications and then, upon finding that experiments do not have logically impelled conclusions, to ascribe

[1] A collection of this work may be found in Achinstein and Hannaway 1985. For a good selection of interdisciplinary contributions, see Gooding, Pinch, and Schaffer 1989.

the experimentalists' beliefs entirely to 'interests'" (p. 11). He suggests that we look instead at how experimentalists actually reason.

Similarly, Allan Franklin (1986, 1990) finds in experimental practice the key to combating doubts about the power of empirical evidence in science. He puts forward what he calls an "evidence model" of science—"that when questions of theory choice, confirmation, or refutation are raised they are answered on the basis of valid experimental evidence" (1990, 2)—in contrast to the view that science is merely a social construction.

An evaluation of the New Experimentalists' success must distinguish between their having provided us sticks with which to beat the social constructivists and their having advanced solutions to philosophical problems that persist, even granting that evidence provides an objective constraint in science.

2. Experiment May Have a Life of Its Own

This slogan, from Hacking 1983, 1992a, and 1992b points to several New Experimentalist subthemes, and can be read in three ways, each in keeping with the position I developed in chapter 2.

Topical hypotheses. The first sense, which Hacking (1983, 160) emphasizes, concerns the aims of experiment. In particular, he and others recognize that a major aim of experiment is to learn things without any intention of testing some theory. In a more recent work, Hacking calls the kinds of claims that experiment investigates "topical hypotheses"—like topical creams—in contrast to deeply penetrating theories. Hacking (1992a) claims that

> it is a virtue of recent philosophy of science that it has increasingly come to acknowledge that most of the intellectual work of the theoretical sciences is conducted at [the level of *topical* hypotheses] rather than in the rarefied gas of systematic theory. (P.45)

The New Experimentalists have led in this recognition. Galison (1987) likewise emphasizes that the

> experimentalists' real concern is not with global changes of world view. In the laboratory the scientist wants to find local methods to eliminate or at least quantify backgrounds, to understand where the signal is being lost, and to correct systematic errors. (P. 245)

The parallels with our recasting of Kuhnian normal science in the last chapter are clear.

Theory-independent warrant for data. A second reading of the slogan refers to the justification of experimental evidence—that a theory-independent warrant is often available. More precisely, the thesis is that experimental evidence need not be theory laden in any way that invalidates its various roles in grounding experimental arguments. Granting that experimental data are not just given unproblematically, the position is that coming to accept experimental data can be based on experimental processes and arguments whose reliability is independently demonstrated. Some have especially stressed the independent grounding afforded by knowledge of instruments; others stress the weight of certain experimental activities, such as manipulation. The associated argument, in each case, falls under what I am calling *exemplary* or *canonical arguments* for learning from error.

Experimental knowledge remains. This reading leads directly to the third gloss of the slogan about experiment having a life of its own. It concerns the continuity and growth of experimental knowledge. In particular, the New Experimentalists observe, experimental knowledge remains despite theory change. Says Galison, "Experimental conclusions have a stubbornness not easily canceled by theory change" (1987, 259). This cuts against the view that holders of different theories necessarily construe evidence in incommensurable or biased fashions. We saw in chapter 2 that the criteria of good normal science lead to just the kind of reliable experimental knowledge that remains through global theory change. These norms, I argued, themselves belie the position Kuhn takes on revolutionary science, where everything allegedly changes.

Continuity at the level of experimental knowledge also has ramifications for the question of scientific progress. It points to a crucial kind of progress that is overlooked when measures of progress are sought only in terms of an improvement in theories or other larger units. Experimental knowledge grows, as do the tools for its acquisition, including instrumentation, manipulation, computation, and most broadly, argumentation.[2] Giere and Hacking have especially stressed how this sort of progress is indicated when an entity or process becomes so well understood that it can be used to investigate other objects and processes (e.g., Giere 1988, 140). We can happily accept what

[2] Ackermann (1985) and others stress progress through instrumentation, but there is also progress by means of a whole host of strategies for obtaining experimental knowledge. That is why I take progress through experimental argumentation to be the broadest category of experimental progress.

these authors say about experimental progress while remaining agnostic about what this kind of progress might or might not show about the philosophical doctrine of scientific realism.

3. What Experimentalists Find: Emphasis on Local Discrimination of Error

A third general theme of New Experimentalist work concerns the particular types of tasks that scientists engage in when one turns to the processes of obtaining, modeling, and learning from experimental data: checking instruments, ruling out extraneous factors, getting accuracy estimates, distinguishing real effect from artifact. In short, they are engaged in the manifold local tasks that may be seen as estimating, distinguishing, and ruling out various errors (in our broad sense).

"How do experiments end?" (as in the title of Galison's book) asks "When do experimentalists stake their claim on the reality of an effect? When do they assert that the counter's pulse or the spike in a graph is more than an artifact of the apparatus or environment?" (Galison 1987, 4). The answer, in a nutshell, is only after they have sufficiently ruled out or "subtracted out" various backgrounds that could be responsible for an effect. "As the artistic tale suggests," Galison continues, "the task of removing the background is not ancillary to identifying the foreground—*the two tasks are one and the same*" (p. 256), and the rest of his book explores the vast and often years-long effort to conduct and resolve debates over background.

3.2 What Might an Epistemology of Experiment Be?

Now to build upon the three themes from the New Experimentalist work, which may be listed as follows:

1. Understanding the role of experiment is the key to circumventing doubts about the objectivity of observation.

2. Experiment has a life of its own apart from high level theorizing (pointing to a local yet crucially important type of progress).

3. The cornerstone of experimental knowledge is the ability to discriminate backgrounds: signal from noise, real effect from artifact, and so on.

In pressing these themes, many philosophers of science sense that the New Experimentalists have opened a new and promising avenue within which to grapple with the challenges they face. Less clear is whether the new attention to experiment has paid off in advancing solutions to problems. Nor is it even clear that they have demarcated a program for working out a philosophy or epistemology of experiment.

The New Experimentalist work seems to agree on certain central questions of a philosophy or epistemology of experiment: how to establish well-grounded observational data, how to use data to find out about experimental processes, and how this knowledge bears on revising and appraising hypotheses and theories. Satisfactory answers to these questions would speak to many key problems with which philosophers of science wrestle, but the New Experimentalist work has not yet issued an account of experimental data adequate to the task.

Experimental activities do offer especially powerful grounds for arriving at data and distinguishing real effects from artifacts, but what are these grounds and why are they so powerful? These are core questions of this book and can be answered adequately only one step at a time.

As a first step we can ask, What is the structure of the argument for arriving at this knowledge? My answer is the one sketched in chapter 1: it follows the pattern of *an argument from error* or *learning from error.* The overarching structure of the argument is guided by the following thesis:

> It is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests) having a high probability of detecting the error if (and only if[3]) it exists nevertheless fails to do so, but instead produces results that accord well with the *absence* of the error.

Such a procedure of inquiry is highly capable of severely probing for errors—let us call it 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.

Alternatively, the argument from error can be described in terms of a test of a hypothesis, *H,* that a given error is absent. The evidence indicates the correctness of hypothesis *H,* when *H* passes a severe test—one with a high probability of failing *H,* if *H* is false. An analogous argument is used to infer the presence of an error.

With this conjecture in hand, let us return to the New Experimentalists. I believe that their offerings are the most interesting and illuminating for the epistemologist of science when they reveal (unwittingly or not) strategies for arriving at especially strong experimental argu-

[3] The "only if" is already accommodated by the requirement that failing to detect means the result is probable assuming the error is absent. The extreme case would be if the result is entailed by the absence of the error. This thesis is the informal side of the more formal definition of passing a severe test in chapter 6.

ments; and when this is so, I maintain, it is because they are describing ways to arrive at procedures with the capacity to probe severely and learn from errors. The following two examples, from Galison 1987 and Hacking 1983, though sufficient, could easily be multiplied.

Arguments for Real Effects

Galison

> The consistency of different data-analysis procedures can persuade the high-energy physicist that a real effect is present. A similar implicit argument occurs in smaller-scale physics. On the laboratory bench the experimenter can easily vary experimental conditions; when the data remain consistent, the experimentalist believes the effect is no fluke. (1987, 219)

What is the rationale for being thus persuaded that the effect is real? The next sentence contains the clue:

> In both cases, large- and small-scale work, the underlying assumption is the same: under sufficient variation any artifact ought to reveal itself by causing a discrepancy between the different "subexperiments." (Ibid.)

Although several main strategies experimentalists use lie scattered through Galison's narratives, he does not explicitly propose a general epistemological rationale for the inferences reached. The argument from error supplies one.

How do these cases fit the pattern of my argument from error? The evidence is the consistency of results over diverse experiments, and what is learned or inferred is "that it is no fluke." Why does the evidence warrant the no-fluke hypothesis? Because were it a fluke, it would almost surely have been revealed in one of the deliberately varied "subexperiments." 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 an artifact by *not* yielding such consistent results. Never mind just now how to justify the probative power (severity) of the procedure. For the moment, we are just extracting a core type of argument offered by the New Experimentalists. And the pattern of the overall argument is that of my argument from error.

Galison's discussion reveals a further insight: whether it is possible to vary background factors or use data analysis to argue "as if" they are varied, the aim is the same—to argue from the consistency of results to rule out its being due to an artifact.

Hacking. An analysis of Hacking's "argument from coincidence" reveals the same pattern, although Hacking focuses on cases where it is possible to vary backgrounds by way of literal manipulation.

Hacking asks, What convinces someone that an effect is real? Low-powered electron microscopy reveals small dots in red blood platelets, called dense bodies. Are they merely artifacts of the electron microscope?

> One test is obvious: can one see these selfsame bodies using quite different physical techniques? . . . In the fluorescence micrographs there is exactly the same arrangement of grid, general cell structure and of the "bodies" seen in the electron micrograph. It is inferred that the bodies are not an artifact of the electron microscope. . . . It would be a preposterous coincidence if, time and again, two completely different physical processes produced identical visual configurations which were, however, artifacts of the physical processes rather than real structures in the cell. (Hacking 1983, 200–201)

Two things should again be noted: First, the aim is the local one, to distinguish artifacts from real objects or effects—something that can be pursued even without a theory about the entities or effects in question. Second, Hacking's argument from coincidence is an example of sustaining an argument from error. The error of concern is to take as real structure something that is merely an artifact. The evidence is the identical configurations produced by completely different physical processes. Such evidence would be extremely unlikely if the evidence were due to "artifacts of the physical processes rather than real structures in the cell" (ibid.).

As before, much more needs to be said to justify this experimental argument, and Hacking goes on to do that; for example, he stresses that we made the grid, we know all about these grids, and so on. But the present concern is the *pattern* of the argument, and it goes like this: "If you can see the same fundamental features of structure using several different physical systems, you have excellent reason for saying, 'that's real' rather than, 'that's an artifact'" (Hacking 1983, 204).[4] It is not merely the *improbability* of all the instruments and techniques conspiring to make all of the evidence appear as if the effect were real. Rather, to paraphrase Hacking again, it is the fact that it would be akin to invoking a Cartesian demon to suppose such a conspiracy.

Hacking tends to emphasize the special evidential weight afforded by performing certain experimental *activities* rather than what is af-

[4] It is important to distinguish carefully between "real" as it is understood here, namely, as genuine or systematic, and as it is understood by various realisms.

forded by certain kinds of *arguments*, saying that "no one actually produces this 'argument from coincidence' in real life" (1983, 201). Not so. Unless it is obvious that such an argument *could be given*, experimental practitioners produce such arguments all the time. In any event, as seekers of a philosophy of experiment we need to articulate the argument if we are to carry out its tasks. The tasks require us to get at the structure of experimental activities and at the epistemological rationale for inferences based on the results of such activities. Most important, understanding the argument is essential in order to justify inferences where the best one can do is simulate, mimic, or otherwise *argue as if* certain experimental activities (e.g., literal manipulation or variation) had occurred. Experimental arguments, I suggest, often serve as surrogates for actual experiments; they may be seen as experiments "done on paper."

Other examples from the work of other New Experimentalists (e.g., conservation of parity), as well as the work of other philosophers of science, lend themselves to an analogous treatment, but these two will suffice. In each case, what are needed are tools for arriving at, communicating, and justifying experimental arguments, and for using the results of one argument as input into others. Those aspects of experiment that have the most to offer in developing such tools are still largely untapped, however, which explains, I think, why the New Experimentalists have come up short. These aspects cover the designing, modeling, and analyzing of experiments—activities that receive structure by means of standard statistical methods and arguments.

Put Your Epistemology of Experiment at the Level of Experiment

In rejecting old-style accounts of confirmation as the wrong way to go, the New Experimentalists seem dubious about the value of utilizing statistical ideas to construct a general account of experimental inference. Theories of confirmation, inductive inference, and testing were born in a theory-dominated philosophy of science, and this is what they wish to move away from. It is not just that the New Experimentalists want to sidestep the philosophical paradoxes and difficulties that plagued formal attempts at inductive logics. The complexities and context dependencies of actual experimental practice just seem recalcitrant to the kind of uniform treatment dreamt of by philosophers of induction. And since it is felt that overlooking these complexities is precisely what led to many of the problems that the New Experimentalists hope to resolve, it is natural to find them skeptical of the value of general accounts of scientific inference.

The typical features of what may be called "theory-dominated" accounts of confirmation or testing are these: (1) the philosophical work begins with data or evidence statements already in hand; (2) the account seeks to provide uniform rules for relating evidence (or evidence statements) to any theory or conclusion (or decision) of interest; and (3) as a consequence of (1) and (2), the account functions largely as a way to *reconstruct* a scientific inference or decision, rather than giving us tools scientists actually use or even a way to model the tools actually used.

The New Experimentalists are right to despair of accounts that kick in only after sufficiently sharp statements of evidence and hypotheses are in hand. Galison is right to doubt that it is productive to search for "an after-the-fact reconstruction based on an inductive logic" (1987, 3). Where the New Experimentalists shortchange themselves is in playing down the use of local statistical methods at the experimental level—the very level they exhort us to focus on.[5] The experimental narratives themselves are chock-full of applications of standard statistical methods, methods developed by Fisher, Neyman and Pearson, and others. Despite the alleged commitment to the actual practices of science, however, there is no attempt to explicate these statistical practices on the scientists' own terms. Ironically, where there is an attempt to employ statistical methods to erect an epistemology of experiment, the New Experimentalists revert to the theory-dominated philosophies of decision and inference. A good example is Allan Franklin's appeal to the Bayesian Way in attempting to erect a philosophy of experiment.

The conglomeration of methods and models from standard error statistics, error analysis, experimental design, and cognate methods, I will argue, is the place to look for forward-looking procedures that serve to obtain data in the first place, and that are apt even with only vague preliminary questions in hand. If what I want are tools for discriminating signals from noise, ruling out artifacts, and so on, then I really need tools for doing that. And these tools must be applicable with the kinds of information scientists actually have. At the same time, these tools can provide the needed structure for the practices

[5] Hacking's recent work often comes close to what I have in mind, but he is reluctant to worry about the Bayes versus non-Bayes controversy in philosophy of statistics. "It is true that different schools will give you different advice about how to design experiments, but for any given body of data they agree almost everywhere" (Hacking 1992b, 153). When it comes to the use of statistical ideas for a general philosophy of experiment, the divergent recommendations regarding experimental design are crucial. It is as regards the uses of a theory of statistics *for philosophy of science*—the uses that interest me—that the debates in philosophy of statistics matter.

given a central place by the New Experimentalists. Before turning to standard error statistics we need to consider why the Bayesian Way fails to serve the ends in view.

3.3 The Bayesian Way

> I take a natural and realistic view of science to allow for the acceptance of corrigible statements, both in the form of data and in the form of laws and hypotheses. . . . It is hard to see what motivates the Bayesian who wants to replace the fabric of science, already complicated enough, with a vastly more complicated representation in which each statement of science is accompanied by its probability, for each of us. (Kyburg 1993, 149)

By the "Bayesian Way" I mean the way or ways in which a certain mathematical theory of statistical inference—Bayesian inference—is used in philosophy of science. My criticism is of its uses regarding philosophical problems of scientific inference and hypothesis testing as distinct from its use in certain statistical contexts (where the ingredients it requires, e.g., prior probabilities, are unproblematic or less problematic) and in personal decision-making contexts. It is not that the Bayesian approach is free of problems in these arenas—ongoing controversies are ever present among statisticians and philosophers of statistics. But the problems of central interest to the philosopher of the epistemology of experiment are those that concern the Bayesian Way in philosophy of science, specifically, problems of scientific inference and methodology.[6]

There is another reason to focus on the Bayesian Way in philosophy of science: it is here that the deepest and most philosophically relevant distinctions between Bayesian and non-Bayesian ideas emerge. For a set of well-defined statistical problems, and for given sets of data, Bayesian and non-Bayesian inferences may be found to formally agree—despite differences in interpretation and rationale. When it comes to using statistical methods in philosophy of science,

[6] In making this qualification I mean to signal that I recognize the relevance of decision theory to philosophy of science generally. However, the set of philosophical problems for which decision theory is most applicable are distinct from those of scientific inference. It is true that decision theory (Bayesian and non-Bayesian) has also been used to model rational scientific agents. However, I do not find those models useful for my particular purpose, which has to do with identifying rational methods rather than rational agents or rational actions. It would take us too far afield to consider those models here.

differences in experimental design, interpretation, and rationales are all-important.

There are three main ways in which a mathematical theory of probabilistic or statistical inference can be used in philosophy of science:

1. *A way to model scientific inference.* The aim may be to model or represent certain activities in science, such as acquiring data, making inferences or decisions, and confirming or testing hypotheses or theories. The intention may be to capture either actual or rational ways to carry out these activities.

2. *A way to solve problems in philosophy of science.* The aim may be to help solve philosophical problems concerning scientific inference and observation (e.g., objectivity of observation, underdetermination, Duhem's problem).

3. *A way to perform a metamethodological critique.* It can be used to scrutinize methodological principles (according special weight to "novel" facts) or to critique the rationality of scientific episodes (metamethodology).

There are other ways of using a theory of statistics, but the above are the most relevant to the epistemological issues before us. The Bayesian Way has in fact been put to all these uses, and many imagine that it is the only plausible way of using ideas from mathematical statistics to broach these concerns in the philosophy of science. Indeed, its adherents often tout their approach as the only account of inference we will ever need, and some unblushingly declare the Bayesian Way to be the route toward solving all problems of scientific inference and methodology. I do not agree.

Although the Bayesian literature is long and technical, explaining why the Bayesian Way is inadequate for each of the three aims requires little or no technical statistics. Such an explication seems to me to be of pressing importance. Keeping the Bayesian philosophy of science shrouded in mathematical complexity has led to its work going on largely divorced from other approaches in philosophy of science. Philosophers of science who do consult philosophers of statistics get the impression that anything but Bayesian statistics is discredited. Thus important aspects of scientific practice are misunderstood or overlooked by philosophers of science because these practices directly reflect non-Bayesian principles and methods that are widespread in science.

It may seem surprising, given the current climate in philosophy of science, to find philosophers (still) declaring invalid a standard set of experimental methods rather than trying to understand or explain why scientists evidently (still) find them so useful. I think it is surprising. Is there something special about the philosophy of experimental inference that places it outside the newer naturalistic attitudes? By and large, Bayesian statisticians proceed as if there were. Colin Howson and Peter Urbach (1989) charge "that one cannot derive scientifically significant conclusions from the type of information which the Fisher and the Neyman-Pearson theories regard as adequate" (p. 130), despite the fact that for decades scientists and statisticians have made it clear they think otherwise. Nor is their position an isolated case. Howson and Urbach are simply the most recent advocates of the strict Bayesian line of argument worked out by fathers of Bayesianism such as Bruno De Finetti, I. J. Good, Denis Lindley, and L. J. Savage. To their credit, Howson and Urbach attempt to apply the Bayesian Way to current challenges in philosophy of science, and so are useful to our project.

Granted, the majority of Bayesians seem to want to occupy a position less strict than that espoused by Howson and Urbach, although they are not entirely clear about what this means. What is clear is that thus far none of the middle-of-the-road, fallen, or otherwise better-behaved Bayesians have promoted the battery of non-Bayesian methods as the basis for an epistemology of experiment. I hope to encourage a change in that direction. I do not believe that an adequate philosophy of experiment can afford to be at odds with statistical practice in science.

The Focus of My Critique: Bayesian Subjectivism

My critique of Bayesianism in this chapter will focus on the first two ways of using an account of statistical inference in philosophy of science—to model scientific inference and to solve philosophical problems about scientific inference.[7] Simply, the Bayesian tools do not tell us what we want to know in science. What we seek are ampliative rules for generating and analyzing data and for using data to learn about experimental processes in a reliable and intersubjective manner. The kinds of tools needed to do this are crucially different from those the Bayesians supply.

The shortcomings of the Bayesian Way for the first two aims bear directly on its appropriateness for the third aim—using Bayesian prin-

[7] Let me confess right off that I will give short shrift to many important technical qualifications, historical footnotes, and significant mathematical developments. No doubt some will take me to task for this, and I apologize. For my purposes, I believe, it is of greater importance to get at the main issues in as nontechnical and noncumbersome a manner as possible.

ciples in a metamethodological critique. (Bayesian critiques of non-Bayesian principles and methods will be addressed in later chapters.)

My immediate target is the version or versions of Bayesianism routinely appealed to by philosophers of the Bayesian Way: the standard subjective Bayesian account (with a few exceptions to be noted).[8] To keep the discussion informal I shall proceed concentrically, going once over the main issues, then again more deeply—rotating all the while among the New Experimentalist program, philosophy of statistics, and philosophy of science. In later loops (and later chapters) some of the more formal notions will fall into place. Although proceeding thus means building an argument piecemeal throughout this book, I can make several of my main points now by looking at how the subjective Bayesians, or Personalists, themselves view the task of an account of statistical or inductive inference.

Evidential-Relationship versus Testing Approaches

In delineating approaches to statistical inference, I find it helpful to distinguish between "evidential-relationship" (E-R) approaches and "testing" approaches. E-R approaches grew naturally from what was traditionally thought to be required by a "logic" of confirmation or induction. They commonly seek quantitative measures of the bearing of evidence on hypotheses. What I call testing approaches, in contrast, focus on finding general methods or procedures of testing with certain good properties.

For now, the distinction between E-R and testing approaches may be regarded as simply a way to help put into perspective the different accounts that have been developed. Only later will this descriptive difference be seen to correspond to more fundamental, epistemological ones. A main way to contrast the two approaches is by means of their quantitative measures. The quantities in E-R approaches are probabilities or other measures (of support or credibility) assigned to hypotheses. In contrast, testing approaches do not assign probabilities to hypotheses. The quantities and principles in testing approaches refer only to properties of methods, for example, of testing or of estimation procedures. One example is the probability that a given procedure of testing would reject a null hypothesis erroneously—an error probability. Another is our notion of a severe testing process.

Bayesian inference is an E-R approach, as I am using that term,

[8] Many of my remarks here and in chapter 10 also apply to so-called objective Bayesians, e.g., Roger Rosenkrantz. For an excellent critical discussion of objective Bayesianism, see Seidenfeld 1979b.

while testing approaches include non-Bayesian approaches, for example, Popperian corroboration, Fisherian statistics, and Neyman-Pearson statistics. Under the category of a testing approach, I would also include entirely qualitative non-Bayesian approaches, for example, those of Clark Glymour and John Worrall.[9]

In the Bayesian approach the key E-R measure is that of a probability of a hypothesis relative to given data. Computing such probabilities requires starting out with a probability assignment, and a major source of difficulty has been how to construe these *prior probabilities.* One way has been to construe them as "logical probabilities," a second, as subjective probabilities.

Carnapian Bayesians. The pioneer in developing a complete E-R theory based on logical probability is Rudolf Carnap.[10] The Carnapian Bayesian sought to assign priors by deducing them from the logical structure of a particular first order language. The E-R measure was to hold between two statements, one expressing a hypothesis and the other data, sometimes written as $C(h,e)$. The measure was to reflect, in some sense, the "degree of implication" or confirmation that e affords h. Calculating its value, the basis for Carnapian logics of confirmation, was a formal or syntactical matter, much like deductive logic.

Such logics of confirmation, however, were found to suffer from serious difficulties. The languages were far too restricted for most scientific cases, a problem never wholly overcome. Even where applicable, a deeper problem remained: How can a priori assignments of probability be relevant to what can be expected to actually occur, that is, to reliability? How can they provide what Wesley Salmon calls "a guide to life"? There is the further problem, Carnap showed, of having to pick from a continuum of inductive logics. To restrict the field, Carnap was led to articulate several postulates, but these, at best, seemed to rest on what Carnap called "inductive intuition." Salmon remarks:

[9] I have hardly completely covered all non-Bayesian approaches. Notable non-Bayesian accounts not discussed are those of Glymour, Kyburg, and Levi. A sect of Bayesians who explicitly consider error probabilities, e.g., Seidenfeld, might seem to be anomalous cases. I regard them as more appropriately placed under the category of testing approaches. I return to this in chapter 10.

[10] Carnap 1962, *Logical Foundations of Probability.* See also Carnap's "Replies and Systematic Expositions" in Schilpp's *The Philosophy of Rudolph Carnap* (Schilpp 1963). Wesley Salmon, in many places, clearly and comprehensively discusses the developments of Carnap's work on induction. See, for example, Salmon 1967 and 1988. See also Carnap and Jeffrey 1971.

> Carnap has stated that the ultimate justification of the axioms is inductive intuition. I do not consider this answer an adequate basis for a concept of rationality. Indeed, I think that *every* attempt, including those by Jaako Hintikka and his students, to ground the concept of rational degree of belief in logical probability suffers from the same unacceptable apriorism. (Salmon 1988, 13)

Subjective Bayesians. The subjective Bayesian, instead, views prior probabilities as personal degrees of belief on the part of some individual. Subjective Bayesianism is a natural move for inductive logicians still wanting to keep within the general Carnapian (E-R) tradition of what an inductive logic should look like. By replacing logical with subjective probabilities, it provides an evidential-relationship approach to confirmation without the problems of logical probability. The definition and tasks of inductive logic become altered correspondingly. Take Howson and Urbach 1989:

> Inductive logic—which is how we regard the subjective Bayesian theory—is the theory of inference from some exogenously given data and prior distribution of belief to a posterior distribution. (P. 290)

The prior distribution of belief refers to the degrees of belief an agent has in a hypothesis *H* and its alternatives prior to the data;[11] the posterior (or final) distribution refers to the agent's degree of belief in *H* after some data or evidence statement is accepted. Inductive inference from evidence is a matter of updating one's degree of belief to yield a posterior degree of belief (via Bayes's theorem).[12]

The Bayesian conception of inductive logic reflects a key feature of theory-dominated philosophies of science: an account of inference begins its work only after sufficiently sharp statements of evidence and hypotheses are in hand. But more is required to get a Bayesian inference off the ground. Also necessary are assignments of degrees of belief to an exhaustive set of hypotheses that could explain the evidence. These are the prior probability assignments. (The full-dress Bayesian requires utilities as well, but I leave this to one side.)

Where do the prior probabilities come from? How does one come to accept the evidence? That the Bayesian approach places no restric-

[11] When, as is often the case, the data are known, the prior probability refers to the degree of belief the agent supposes he or she would have if the data were not known. Problems with this occupy us later (chapter 10).

[12] Attempts at interval valued probabilities have been proposed but with mixed success. At any rate, nothing in the present discussion is altered by those approaches.

tions on what can serve as hypotheses and evidence, while an important part of its appealing generality, makes it all the more difficult to answer these questions satisfactorily.

Prior Probabilities: Where From?

Many philosophers would agree with Isaac Levi that "strict Bayesians are legitimately challenged to tell us where they get their numbers" (Levi 1982, 387). In particular, it seems they should tell us how to assign prior probabilities. The subjectivist disagrees. The Bayesian subjectivist typically maintains that

> we are under no obligation to legislate concerning the methods people adopt for assigning prior probabilities. These are supposed merely to characterise their beliefs subject to the sole constraint of consistency with the probability calculus. (Howson and Urbach 1989, 273)

Agents presumably are to discover their degrees of belief by introspection, perhaps by considering the odds they might give if presented with (and required to take?) a series of bets.

But would not such personal opinions be highly unstable, varying not just from person to person, but from moment to moment? That they would, subjectivists accept and expect.

In their classic paper, Edwards, Lindman, and Savage (1963) tell us that the probability of a hypothesis *H*, *P*(*H*) "might be illustrated by the sentence: 'The probability for you, now, that Russia will use a booster rocket bigger than our planned Saturn booster within the next year is .8'" (p. 198). Throughout the introductory text by Richard Savage (L. J.'s brother) "my probability" is quite deliberately used instead of "probability."

Quantitatively expressing the degree of belief "for you now" is quite outside what Bayesian inference officially supplies. Bayesian inference takes it as a given that agents have degrees of belief and assumes that these are expressible as probabilities; its work is to offer a way of fitting your beliefs together coherently. In particular, your beliefs prior to the data should cohere with those posterior to the data by Bayes's theorem (whether you do it by conditionalization or by changing your prior probability assignment[13]).

[13] Not all Bayesians hold the posterior to result from conditionality only. It might be due to a change in prior probability assignment for reasons other than new evidence *e*. Those who violate conditionality have the Bayesian approach doing even less work—it only tells you to be coherent. Nothing in our discussion turns on this qualification, however.

(Much of the technical work by Bayesian philosophers concerns so-called Dutch Book arguments, which come in various forms. These arguments purport to show that if we are rational, we will be coherent in the Bayesian sense. The basic argument is that if it is given that beliefs are expressible as probabilities, then, assuming you must accept every bet you are offered, if your beliefs do not conform to the probability calculus, you are being incoherent and will lose money for sure. In as much as these givens hardly seem to describe the situation in science, as many have argued,[14] we need not accept what such arguments purport to show.)

Personal Consistency versus Scientific Prediction

Bayes's theorem, to be stated shortly, follows from the probability calculus and is unquestioned by critics. What is questioned by critics is the relevance of a certain use of this theorem, namely, for scientific inference. Their question for the subjective Bayesian is whether scientists have prior degrees of belief in the hypotheses they investigate and whether, even if they do, it is desirable to have them figure centrally in learning from data in 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.

Leonard "L. J." Savage, a founder of modern subjective 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; emphasis added). Fittingly, Savage employs the term "personalism" to describe subjective Bayesianism.

But is a personal code of consistency, requiring the quantification of personal opinions, however vague or ill formed, an appropriate basis for scientific inference? Most of the founders of modern statistical theory—Fisher, Neyman, Pearson, and others—said no. Pearson (of Neyman and Pearson) put his rejection this way:

> It seems to me that . . . [even with no additional knowledge] I might quote at intervals widely different Bayesian probabilities for the same set of states, simply because I should be attempting what would be for me impossible and resorting to guesswork. It is difficult to see how the matter could be put to experimental test. (Pearson 1966e, 278)

Stating his position by means of a question, as he was wont to do, Pearson asks:

[14] For an excellent recent discussion, see Baccus, Kyburg, and Thalos 1990.

> Can it really lead to my own clear thinking to put at the very foundation of the mathematical structure used in acquiring knowledge, functions about whose form I have often such imprecise ideas? (Pearson 1966e, 279)

Fisher expressed his rejection of the Bayesian approach far more vehemently (which is not to say that he favored the one erected by Neyman and Pearson, but more on that later). Bayesians, Fisher declared,

> seem forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes. (Fisher 1947, 6–7)

As is evident from this chapter's epigraph, Fisher denied the need for posterior probabilities of hypotheses in science in the first place.

In an earlier generation (late nineteenth century), C. S. Peirce, anticipating the later, non-Bayesian statisticians, similarly criticized the use of subjective probabilities in his day. Considering Peirce will clarify Fisher's claim that Bayesians "seem forced to regard" probability as subjective degrees of belief.

Why the Evidential-Relationship Philosophy Leads to Subjectivism

Peirce, whom I shall look at more closely in chapter 12, is well aware that probabilities of hypotheses are calculable by the doctrine of "inverse probability" (Bayes's theorem). However, Peirce explains,

> this depends upon knowing antecedent probabilities. If these antecedent probabilities were solid statistical facts, like those upon which the insurance business rests, the ordinary precepts and practice [of inverse probability] would be sound. But they are not and cannot be statistical facts. What is the antecedent probability that matter should be composed of atoms? Can we take statistics of a multitude of different universes? . . . All that is attainable are subjective probabilities. (Peirce 2.777)[15]

And subjective probabilities, Peirce continues, "are the source of most of the errors into which man falls, and of all the worst of them" (ibid.).

By "solid statistical facts" Peirce means that they have some clear stochastic or frequentist interpretation. (I discuss my gloss on frequentist statistics in chapter 5.) It makes sense to talk of the relative

[15] All Peirce references are to C. S. Peirce, *Collected Papers*. References are cited by volume and paragraph number. For example, Peirce 2.777 refers to volume 2, paragraph 777.

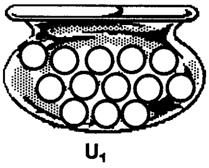

FIGURE 3.1. A single-universe context.

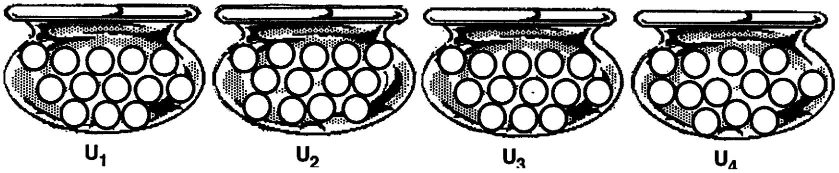

FIGURE 3.2. A multiple-universe context (universes as plenty as blackberries).

frequency of events such as "heads" in a population of coin-tossing experiments but not of the relative frequency of the truth of a hypothesis such as matter is composed of atoms. The probability of a hypothesis would make sense, Peirce goes on to say, only

> if universes were as plenty as blackberries, if we could put a quantity of them in a bag, shake them well up, draw out a sample and examine them to see what proportion of them had one arrangement and what proportion another. (2.684)

Single- versus Multiple-Universe Contexts. Figures 3.1 and 3.2 illustrate the distinction between the situation Peirce regards scientists as being in and one where universes are "as plenty as blackberries." The first situation (fig. 3.1) involves just one universe or one urn. It contains, let us suppose, some fixed proportion of white balls.

The second situation (fig. 3.2) involves some (possibly infinite) number of urns, U_1, U_2, . . . , each with some proportion of white balls. Here universes, represented as urns, are "as plenty as blackberries." Consider a hypothesis about the single-universe context, hypothesis *H:* the proportion of white balls (in this one universe U_1) equals .6. *H* asserts that in a random selection from that urn, the probability of a white ball equals .6. Because there is just this one urn, hypothesis *H* either is or is not true about *it*. *H* is true either 100 percent of the time or 0 percent of the time. The only probabilities that could be assigned to *H* itself are the trivial ones, 1 and 0.

Now consider the second situation, where there are many universes or urns. Hypothesis *H* may be true in some and not in others. If we can imagine reaching in and drawing out one of the universes, like selecting well-shaken blackberries from a bag, then it makes sense to talk about the probability that we will select a universe (urn) in which *H* is true. (A specific example to come.) But the second context is not our situation in science, says Peirce. Since our scientific hypotheses refer to just this one universe that we are in, like it or not, probabilities of such hypotheses cannot be regarded as "solid statistical facts."

By interpreting probabilities as subjective degrees of belief in hypotheses, however, it becomes meaningful to talk about nontrivial assignments of probabilities to hypotheses—even about this one universe. Until a hypothesis is known to be true or false by an agent, the agent may be supposed to have some quantitative assessment, between 0 and 1, of the strength of credibility one feels toward the hypothesis. Believing a certain "Big Bang" hypothesis to be very credible, for instance, you might assign it a degree of belief of .9.

We can now understand why the desire for a posterior probability measure, coupled with a single-universe context (as well as the rejection of logical probabilities), "seems to force" the subjective interpretation of the probability calculus, as Fisher alleged. Fisher, Peirce, Neyman, and Pearson, as well as contemporary frequentists, view the attempt to model quantitatively the strengths of opinion and their changes as useless for science. Thus except for contexts appropriately modeled as in figure 3.2—multiple urn experiments—they view theories of statistics appropriate for the second case as inappropriate for scientific inquiry into our one universe.[16]

An Illustration: A Multiple-Universe Context and Bayes's Theorem

It will be useful to have a very simple example of a context that *can* be modeled as a multiple-universe or multiple-urn context in which Bayes's theorem can be applied. It will also help clarify the notion of conditional probability.

a. Consider a game of chance, *rouge et noire:* You bet on either black or red, (randomly) select a card from the deck, and win if it is of a suit with your color. Let the possible outcomes be either "win" or "lose." Suppose that the probability of a win given that *rouge et noire* is played equals .5. We can abbreviate this sentence using probability notation:

$$P(\text{win} \mid \textit{rouge et noire}) = \tfrac{1}{2}.$$

[16] Some may regard Fisher's "fiducial probabilities" as falling outside this delineation, and they may be correct. This is no doubt bound up with the reason that such probabilities lead to inconsistencies.

For our purposes, conditional probability need not be technically explored. Grasp it by reading whatever comes after the "given bar" (|) as announcing the specific type of experiment, condition, or hypothesis to which you are restricted in considering the probability of the outcome of interest. With *rouge et noire* we are asserting that "the probability of winning, given that the experiment is a *rouge et noire* experiment, is one-half."

b. Now consider a second game, that of betting on 1 of 36 numbers in roulette (assume that no 0 or 00 outcomes are on the wheel). Let the probability of a win, given that the second game is played, equal $1/36$. We can write this as

$P(\text{win} \mid \text{single-number game}) = 1/36$.

c. Now consider a third game, a sort of second-order game. A fair coin is tossed to decide whether to play the first or second game above. Say that "heads" results in *rouge et noir* being played, "tails," in the single-number roulette game. Then, with probability $1/2$, *rouge et noire* is played, and with probability $1/2$ the single-number game is played. Games *a* and *b* are like blackberries that we shake up in a bag and draw from. (Never mind why anyone would play this game!)

What has happened in the third game is that the game to be played *is itself* an outcome of a game of chance. That is, there are two outcomes: "*rouge et noire* is played" and "single-number roulette is played"—where it is given that these are the only two possibilities. We can write these as two "hypotheses":

H_1: *rouge et noire* is played.

H_2: single-number roulette is played.

Notice that here we have stipulated that the truth of these two hypotheses is determined by the outcome of a game of chance. Each hypothesis is true with probability $1/2$. That is, the context has two blackberries, H_1 and H_2, and each has equal chance of being drawn from the bag. Thus, we have two (perfectly objective) *unconditional* probabilities:

$P(H_1)$ (i.e., the probability that *rouge et noire* is played)

and

$P(H_2)$ (i.e., the probability that single-number roulette is played).

Further, we know the values of these two unconditional probabilities, because we have stipulated that they are each $1/2$.

$P(H_1) = P(H_2) = 1/2$.

d. Now imagine that you are told the following: a woman who has gone through the game in (*c*), and played whatever game it selected for her, has won. What might be inferred about whether she won through playing *rouge et noire* (H_1 is true) or through single-number roulette (H_2 is true)? The prior (unconditional) probability in each is ½; but with this new information—the result was a win—we can update this probability and calculate the probability that the game played was *rouge et noire* given that it yielded a win. That is, we can calculate the (posterior) conditional probability

$$P(H_1 \mid \text{win}).$$

(We can likewise calculate $P[H_2 \mid \text{win}]$, but let us just do the first.)

The formula for this updating is *Bayes's theorem,* and in this case even one who insists on objective probabilities can use it. It just follows from the definition of conditional probability.[17]

$$P(H_1 \mid \text{win}) = \frac{P(\text{win} \mid H_1)\, P(H_1)}{P(\text{win} \mid H_1)P(H_1) + P(\text{win} \mid H_2)\, P(H_2)}.$$

Here, the needed probabilities for the computation are given. The prior probabilities are given, and from (*a*) and (*b*) we have

$$P(\text{win} \mid H_1) = 1/2, \text{ and}$$
$$P(\text{win} \mid H_2) = 1/36.$$

The reader may want to calculate $P(H_1 \mid \text{win})$. The answer is 18/19. So the evidence of a win gives a Bayesian "confirmation" of hypothesis H_1: the posterior probability exceeds the prior probability.

Bayes's Theorem

We can generalize this result. For an exhaustive set of disjoint hypotheses, $H_1, H_2, \ldots, H_n$, whose probabilities are not zero, and outcome e where $P(e) > 0$:

[17] In general,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

So we have

$$P(H_1 \mid \text{win}) = \frac{P(H_1 \text{ and win})}{P(\text{win})},$$

and

$$P(\text{win}) = P(\text{win and } H_1) + P(\text{win and } H_2).$$

$$P(H_1 \mid e) = \frac{P(e \mid H_1)\, P(H_1)}{P(e \mid H_1)\, P(H_1) + P(e \mid H_2)\, P(H_2) + \ldots + P(e \mid H_n)\, P(H_n)}.$$

A Role for Opinions?

In a case like the above illustration, the truth of a hypothesis can be seen as an outcome of an experimental process, and it makes sense to talk about the probability of that outcome in the usual frequentist sense. Since it makes sense to talk about the probability of the outcome, it makes sense, in this special kind of case, to talk about the probability of the hypothesis being true. In such cases there is no philosophical problem with calculating the posterior probabilities using Bayes's theorem. *Except* for such contexts, however, the prior probabilities of the hypotheses are problematic. Given that logical probabilities will not do, the only thing left is subjective probabilities. For many, these are unwelcome in scientific inquiry. Not to subjectivists.

Subjectivists or personalists, by contrast, seem only too happy to announce, as Savage (1964) puts it, that

> the Bayesian outlook reinstates opinion in statistics—in the guise of the personal probabilities of events. (P. 178)

and that

> the concept of personal probability . . . seems to those of us who have worked with it an excellent model for the concept of opinion. (P. 182)

But whether personal probability succeeds well or badly in modeling opinion—something that is itself open to question—is beside the point for those who, like Peirce, Fisher, Neyman, and Pearson, see this kind of reliance on opinion as entirely wrongheaded for scientific inference. Knowledge of the world, many think, is best promoted by *excluding* so far as possible personal opinions, preferences, and biases. The arguments of Peirce in his day, and of Fisher, and Neyman and Pearson in theirs, remain the major grounds for rejecting the Bayesian approach in science.

Bayesians will object that it is impossible to exclude opinions, that at least the Bayesian brings them out rather than "sweeping them under the carpet," to paraphrase I. J. Good (1976). The charge of subjectivity leveled at non-Bayesian statistics will occupy us later (e.g., in chapters 5 and 11). I will argue that the type of arbitrariness in non-Bayesian error statistics is very different from that of subjective probabilities. The subjectivity of personalism, as Henry Kyburg (1993) has aptly put it, is particularly pernicious.

The Pernicious Subjectivity of Prior Probabilities

That scientists regularly start out with differing opinions in hypotheses is something the subjective Bayesian accepts and expects. Consequently, Bayesian consistency instructs agents to reach different posterior degrees of belief, even on the very same experimental evidence. What should be done in the face of such disagreement? Is there a way to tell who is right? Denis Lindley, also a father of modern Bayesianism, has this to say:

> I am often asked if the method gives the *right* answer: or, more particularly, how do you know if you have got the *right* prior. My reply is that I don't know what is meant by "right" in this context. The Bayesian theory is about *coherence,* not about right or wrong. (Lindley 1976, 359)

It is understandable that Lindley wonders what "right" can mean in the personalist context, for there is no reason to suppose that there is a correct degree of belief to hold. My opinions are my opinions and your opinions are yours. Without some way to criticize prior degrees of belief, it is hard to see how any criticism of your opinion can be warranted. If "right" lacks meaning, how can I say that you are in error? This leads to Kyburg's charge:

> This is almost a touchstone of objectivity: the possibility of error. There is no way I can be in error in my prior distribution for μ—unless I make a logical error—. . . . It is that very fact that makes this prior distribution perniciously subjective. It represents an assumption that has consequences, but cannot be corrected by criticism or further evidence. (Kyburg 1993, 147)

Of course one can change it. Kyburg's point is that even when my degree of belief changes on new evidence, it in no way shows my previous degree of belief to have been mistaken.

The subjectivity of the subjective Bayesian Way presents a major obstacle to its serving as an adequate model for scientific practice. Being right may be meaningless for a personalist, but in scientific contexts being right, avoiding specific errors, is generally well understood. Even where uncertainty exists, this understanding at least guides practitioners toward making progress in settling disagreements. And it guides them toward doing something right now, with the kind of evidence they can realistically obtain.

Swamping Out of Priors

The problem of accounting for consensus is not alleviated by the often heard promise that with sufficient additional evidence differences in prior probability are washed away. For one thing, these "washout theorems" assume that agents assign nonzero priors to the same set of hypotheses as well as agree on the other entries in the Bayesian algorithm. For another, they assume statistical hypotheses, while the Bayesian Way is intended to hold for any type of hypothesis. While some of these assumptions may be relaxed, the results about convergence are far less impressive.

Many excellent critical discussions of these points can be found in the literature.[18] Committed Bayesians will direct me to so-and-so's new theorem that extends convergence results. But these results, however mathematically interesting, are of no help with our problem. The real problem is not that convergence results hold only for very special circumstances; even where they hold they are beside the point. The possibility of eventual convergence of belief is irrelevant to the day-to-day problem of evaluating the evidential bearing of data in science.

Imagine two scientists reporting degrees of belief in *H* of .9 and .1, respectively. Would they find it helpful to know that with some amount of additional data their degree of belief assignments would differ by no more than a given amount? Would they not instead be inclined to dismiss reports of degrees of belief as irrelevant for evaluating evidence in science?[19]

John Earman (1992), despite his valiant efforts to combat the problems of the Bayesian Way, despairs of grounding objectivity via washout theorems:

> Scientists often agree that a particular bit of evidence supports one theory better than another or that a particular theory is better supported by one experimental finding than another. . . . What happens in the long or the short run when additional pieces of evidence are added is irrelevant to the explanation of shared judgments about the evidential value of present evidence. (P. 149)

What, then, explains the consensus about present evidence? Is the choice really, as Earman's title states, "Bayes or Bust"? I see this as a false choice. Science is not a bust. Yet scientists regularly settle or at

[18] See, for example, Earman 1992 and Kyburg 1993. In the case where hypotheses are statistical and outcomes are independent and identically distributed, it is unexceptional that convergence can be expected. It is hard to imagine any theory of statistical inference not having such an asymptotic result for that special case (it follows from the laws of large numbers, chapter 5).

[19] What is more ,the tables can be turned on the washout claims. As Kyburg (1993, 146) shows, for any body of evidence there are prior probabilities in a hypothesis *H* that, while nonextreme, will result in the two scientists having posterior probabilities in *H* that *differ* by as much as one wants.

least make progress with disputes about the import of evidence, and they do so with arguments and analyses based on non-Bayesian principles. The question of how to understand the evidence, in the jargon of chapter 2's Kuhnian analysis, regularly gives rise to a "normal research problem." It is tackled by reliable testing of low-level hypotheses about error.

Making Subjectivism Unimpeachably Objective

Howson and Urbach, staunch defenders of the subjective Bayesian faith, are unfazed by the limited value of the washout theorems, declaring them unnecessary to counter the charge of subjectivity in the first place. The charge, they claim, "is quite misconceived. It arises from a widespread failure to see the subjective Bayesian theory for what it is, a theory of inference. And as such, it is unimpeachably objective: though its subject matter, degrees of belief, is subjective, the rules of consistency imposed on them are not at all" (Howson and Urbach 1989, 290).

Howson and Urbach press an analogy with deductive logic. Just as deductive logic concerns theories of valid inferences from premises to conclusions where the truth of the premises is unknown, inductive logic concerns inferences from premises to some quantitative measure on the conclusion where the truth of the premises is unknown.

When Bayesians talk this way, they reveal just how deeply disparate their view of inductive inference is from what is sought by an account of ampliative inference or experimental learning. Although most Bayesians would not go as far as Howson and Urbach in calling the Bayesian approach "unimpeachably objective," all seem to endorse their analogy between inductive and deductive logic. As Kyburg (1993) has put it, neo-Bayesianism is "yet another effort to convert induction to deduction" (p. 150) in the form of a deductive calculus of probabilities.

A Fundamental Difference in Aims

This fundamental difference in their views of what an account of scientific inference should do has played too little of a role in the Bayes–non-Bayes controversy. Once we recognize that there is a big difference between the goals of a "deductive inductive" inference and what we seek from an ampliative account, we can agree to disagree with Bayesians on the goals of an account of scientific inference. This recognition has two consequences:

First it explains why Bayesian criticisms of non-Bayesian (standard error) statistics cut no ice with non-Bayesians. Such criticisms tend to show only that the latter fail to pass muster on Bayesian grounds. (Examples will occupy us later.) It is true that standard error statistics is "incoherent" according to the Bayesian definition. But Bayesian coherence is of no moment to error statisticians. At a 1970 conference on the foundations of statistics at the University of Waterloo, the statistician Irwin Bross put it bluntly:

> I want to take this opportunity to flatly repudiate the Principle of Coherence which, as I see it, has very little relevance to the statistical inference that is used in the sciences. . . . While we do want to be coherent in ordinary language, it is not necessary for us to be coherent in a jargon that we don't want to use anyway—say the jargon of L. J. Savage or Professor Lindley. (Bross 1971, 448)

This is not to say that all Bayesian criticisms of error statistics may be just dismissed; I will return to them later.

The second consequence of recognizing the difference in aims is more constructive. Conceding the limited scope of the Bayesian algorithm might free the Bayesian to concede that additional methods are needed, if only to fill out the Bayesian account. We will pursue this possibility as we proceed.

Can Bayesians Accept Evidence?

Perhaps the most obvious place for supplementary methods concerns the data or evidence. For just as with arriving at prior probabilities, the Bayesian response when asked about the grounds for accepting data is that it is not their job:

> The Bayesian theory we are proposing is a theory of inference from data; we say nothing about whether it is correct to accept the data. . . . 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)

The idea that we begin from the "acceptance as true of some evidential statement" is problematic for two reasons.[20] First, the Bayesian Way is to hold for any evidence and hypothesis, not just those associated with a specific statistical model. So the evidential statement *e* will

[20] Nor does Richard Jeffrey's (1965) approach, in which the evidence need not be accepted as certain but is allowed to be merely probable, help us. See, for example, the discussion in Kyburg 1974, 118–22.

often be the type of claim that a theory of ampliative inference should help us to assess—not require us to begin with as given. In applying the Bayesian Way to classic episodes of hypothesis appraisal, for example, statements that are called upon to serve as evidence *e* include "Brownian motion exists" and "The estimated deflection of light is such and such."

This leads to the second problem with beginning with accepting the evidence statement as true—one that is more troubling. The standard Bayesian philosophy, after all, eschews acceptance of hypotheses (unless they have probability one), preferring instead to assign them degrees of belief. But why should it be all right to accept a statement when it plays the role of evidence and not when it plays the role of the hypothesis inferred? Now there are Bayesian accounts of the acceptance of hypotheses, but they do not help with our problem of accepting the evidence to get a Bayesian inference going.

Patrick Maher (1993a) proposes "a conception of acceptance that does not make acceptance irrational from a Bayesian perspective" (p. 154). Maher argues (1993a, 1993b), contrary to the general position of Bayesian philosophers of science, that a Bayesian account requires a theory of acceptance to be applicable to the history of science. For the history of science records the acceptance of claims, not scientists' probability assignments. Maher (1993b) argues that Bayesian philosophers of science, for example, Dorling, Franklin, and Howson, "operate with a tacit theory of acceptance" (p. 163)—one that identifies acceptance with high probability. Maher argues that a more adequate Bayesian theory of acceptance is a decision-theoretic one, where acceptance is a function both of probabilities of hypotheses and (cognitive) utilities. While Maher is to be credited for pointing up the shortcomings in Bayesian analyses of scientific episodes, his approach, as with all Bayesian decision approaches, only adds to the ingredients needed to get a Bayesian inference going.[21]

To return to the problem of accepting evidence claims, a solution could possibly be found by supplementing Bayesian algorithms with some separate—non-Bayesian—account for accepting evidential claims. Although the problems of arriving at prior probabilities and setting out alternative hypotheses (and their likelihoods) would still persist, such a supplement might at least offer reliable grounds for accepting the evidence. Still, this tactic would have at least one serious

[21] It would take me too far afield to discuss the various decision-theoretic accounts, Bayesian and non-Bayesian, in this work. Maher (1993b) provides a good overview from the point of view of theories of acceptance.

drawback for Bayesians: the need for a supplementary account of evidence would belie one of the main selling points of the Bayesian approach—that it provides a single, unified account of scientific inference.

All You Need Is Bayes

Subjective Bayesians, especially leaders in the field, are remarkable for their ability to champion the Bayesian Way as the one true way (its many variants notwithstanding). If we take these Bayesians at their word, it appears that they view the Bayesian approach as the only account of inference (and perhaps decision) that we shall ever need. To solve the fundamental problems of inference and methodology, we need only to continue working out the details of the Bayesian paradigm. Consider Lindley's reasoning at the Waterloo conference:

> Now any decision that depends on the data that is being used in making the inference only requires from the data the posterior distribution. Consequently the problem of inference is effectively solved by stating the posterior distribution. This is the reason why I feel that the basic problem of inference is solved. (Lindley 1971, 436)

Even if it were true that stating the posterior solves the problem of inference, it would not follow that the Bayesian Way solves the problem of inference because it does not give one *the* posterior distribution. It gives, at best, a posterior distribution for a given agent reporting *that agent's* degree of belief at a given time. If we restate Lindley's claim to read that the basic problem of inference is solved by stating a given agent's degree of belief in a hypothesis, then I think we must conclude that Lindley's view of the basic problem of inference differs sharply from what scientists view it to be. Learning of an agent's posterior degree of belief, a scientist, it seems to me, would be interested only in what the evidence was for that belief and whether it was warranted. This calls for intersubjective tools for assessing the evidence and for adjudicating disagreements about hypothesis appraisal. This, subjective posteriors do not provide.

M. S. Bartlett, also a prominent statistician attending the Waterloo conference, had this response for Lindley:

> What does professor Lindley mean when he says that "the proof of the pudding is in the eating"? If he has done the cooking it is not surprising if he finds the pudding palatable, but what is his reply if we say that we do not. If the Bayesian allows some general investigation to check the frequency of errors committed . . . this might be

> set up; but if the criterion is inner coherency, then to me this is not acceptable. (Bartlett 1971, 447).

Bartlett puts his finger on a central point over which the error statistician is at loggerheads with the Bayesian: the former's insistence on checking "the frequency of errors" or on *error probabilities*.[22] The centrality of the notion of error probabilities to non-Bayesian statisticians is why it is apt to call them error statisticians. To get a rough and ready idea of the error frequency check for which Bartlett is asking, imagine that the Bayesian agent reports a posterior degree of belief of .9 in hypothesis *H*. Bartlett, an error statistician, would require some way of checking how often such a high assignment would be expected to occur even if *H* is false. What for Bartlett would be necessary for the palatability of the Bayesian posterior, however, would, for the Bayesian, be quite irrelevant. Bayesian principles, as will be seen, conflict with error probability principles (chapter 10). Moreover, error probabilities call for an objective (frequentist[23]) notion of probability—while Bayesians, at least strict (or, to use Savage's [1964] term, "radical") ones, declare subjective probabilities to be all we need.

> I will confess . . . that I and some other Bayesians hold this [personal probability] to be the only valid concept of probability and, therefore, the only one needed in statistics, physics, or other applications of the idea. (Savage 1964, 183)

Scientists, as we shall shortly see, beg to differ. Scientific practice does not support the position that "all you need is Bayes," but the position held by the founders of non-Bayesian methods in the epigraphs to this chapter: when it comes to science, subjective Bayesianism is not needed at all.

Giere: Scientists Are Not Bayesian Agents

In reality, scientists do not proceed to appraise claims by explicit application of Bayesian methods. They do not, for example, report results by reporting their posterior probability assignments to one hypothesis compared with others—even dyed-in-the-wool Bayesians apparently grant this. Followers of the Bayesian Way do not seem very

[22] Those who have investigated the error probabilities of Bayesian methods have found them to be problematic. See, for example, Giere 1969 and Kempthorne and Folks 1971, 304–7. I return to this in chapter 10.

[23] Some, it is true (e.g., Giere), prefer propensities. I will not take up the problems with propensity definitions, but will argue for the appropriateness of frequentist statistics.

disturbed by this. One retort is that they are modeling only the ideally rational scientist, not the actual one. This type of defense is not very comfortable in the present climate where aprioristic philosophy of science is unwelcome. More modern Bayesians take a different tack. They view the Bayesian approach as a way to reconstruct actual scientific episodes and/or to model scientific judgments at some intuitive level, although it is not clear what the latter might mean.

Ronald Giere argues that empirical studies refute the claim that typical scientists are intuitive Bayesians and thereby count against the value of Bayesian reconstructions of science. Giere, a major (non-Bayesian) player in the philosophy of statistics debates of the 1970s, now declares that "we need not pursue this debate any further, for there is now overwhelming empirical evidence that no Bayesian model fits the thoughts or actions of real scientists" (Giere 1988, 149). But I do not think the debate is settled. To see why not we need to ask, What are these empirical studies?

The empirical studies refer to experiments conducted since the 1960s to assess how well people obey Bayes's theorem. These experiments, such as those performed by Daniel Kahneman, Paul Slovic, and Amos Tversky (1982), reveal substantial deviations from the Bayesian model even in simple cases where the prior probabilities are given, and even with statistically sophisticated subjects.

> Human beings are not naturally Bayesian information processors. And even considerable familiarity with probabilistic models seems not generally sufficient to overcome the natural judgment mechanisms, whatever they might be. (Giere 1988, 153)

Apparent success at Bayesian reconstructions of historical cases, Giere concludes, is mistaken or irrelevant.

> Scientists, as a matter of empirical fact, are not Bayesian agents. Reconstructions of actual scientific episodes along Bayesian lines can at most show that a Bayesian agent would have reached similar conclusions to those in fact reached by actual scientists. Any such reconstruction provides no explanation of what actually happened. (Giere 1988, 157)

Although I agree with the upshot of Giere's remarks, I do not claim that the probability experiments are what vitiate the Bayesian reconstructions. While interesting in their own right, these experiments seem to be the wrong place to look to test whether Bayes's theorem is a good model for scientific inference. Why? Because in these experiments the problem is *set up* to be one in which the task is calculating probabilities (whether of a posterior or of a conjunction of claims, or whatever). The experiments refer to classic games of chance or other setups where the needed probabilities are either given or assumed. The probabilities, moreover, refer to objective frequency calculations, not degrees of belief. Even with fully representative subjects, the results are at most relevant to how well people's intuitive judgments of probability obey the calculus of probabilities. They say nothing about whether scientists are engaged in attempting to assign probabilities to the hypotheses about which they inquire. If, as I, and error statisticians, urge, scientific inference is not a matter of assigning probabilities to hypotheses in the manner in which we would assign probabilities to outcomes of games of chance, then it is irrelevant whether subjects' judgments in these contexts accord well or badly with the probability calculus.

The finding of the probability experiments, that humans violate the probability calculus—*when asked to carry out a probability problem*—also says nothing about the methodology actually used in appraising scientific claims. For this we have to look at the kinds of experimental tools and arguments scientists use. One hardly does justice to inferences in science by describing them merely as violations of the Bayesian account. What one finds is a systematic pattern of statistical reasoning—but of the non-Bayesian sort I call standard error statistics. Familiar applications include the typical methods we hear about every day in reports of polling results and of studies on new drugs, cancer-causing substances, and the like. Contrary to what Giere had hoped, the debates concerning these methods still need to be pursued. Only now they should be pursued by considering actual experimental practice.

Where to look? Most classical cases of theory change are too fossilized to help much, unless detailed accounts of data analysis are available. Two such examples (Brownian motion and eclipse experiments) will be considered later. A rich source of examples of standard error statistics in experimental inference is the New Experimentalist narratives that we began discussing earlier. Our discussion now picks up where we left off in section 3.2.

3.4 The New Experimentalists: Experimental Practice Is Non-Bayesian

Regardless of what one thinks of the Bayesian Way's ability to reconstruct or model learning, looking at the tools actually used in the building up of knowledge reveals a use of probabilistic ideas quite unlike that of an after-trial sum-up of a theory's probable truth. I share the view of Oscar Kempthorne, a student of R. A. Fisher:

> It seems then that a use of "probability" as in "the probability that the theory of relativity is correct" does not really enter at all into the building up of knowledge. (Kempthorne and Folks 1971, 505)

One way to capture how probability considerations are used, I propose, is as tools for sustaining experimental arguments even in the absence of literal control and manipulation. The statistical ideas, as I see them, embody much of what has been learned about how limited information and errors lead us astray, as discussed in chapter 1. Using what has been learned about these mistakes, we have erected a conglomeration of interrelated tools that are good at practically forcing mistakes to show themselves, so to speak.

Galison: Neutral Currents

Let us now turn to the tools used in the trenches and brought to light in experimentalist work. Galison's 1987 work is especially congenial, and all references to him in this section refer to that work. Although Galison is not trying to draw any lessons for statistical philosophy and perhaps *because* he is not, his efforts to get at the arguments used to distinguish genuine effect from artifact effectively reveal the important roles served by error statistics. As Galison remarks, a key characteristic of twentieth-century experimental physics is "how much of the burden of experimental demonstration has shifted to data analysis" (p. 151) to distinguish signal from background. The increasingly central role played by data analysis makes the pronouncements of R. A. Fisher and E. S. Pearson (in the epigraphs to this chapter) as relevant today as in their own time.

I shall follow a portion of Galison's discussion of the discovery of neutral currents, thought to be one of the most significant in twentieth-century physics. By the end of the 1960s, Galison tells us, the "collective wisdom" was that there were no neutral currents. Bubble chamber evidence from many experiments indicated that neutral currents either did not exist or were well suppressed (pp. 164, 174). Soon after, however, from 1971 to 1974, "photographs . . . that at first appeared to be mere curiosities came to be seen as powerful evidence for" their existence (p. 135).

This episode, lasting from 1971 to 1974, occupies one-third of Galison's book, but my focus will be on the one analysis for which he provides the most detailed data. Abstracted from the whole story, this part cannot elucidate either the full theory at stake or the sociological context, but it can answer Galison's key question: "How did the experimentalists themselves come to believe 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?" (p. 136).

Here is the gist of their experimental analysis: Neutral currents are described as those neutrino events without muons. Experimental outcomes are described as muonless or muonful events, and the recorded result is the ratio of the number of muonless and muonful events. (This ratio is an example of what is meant by a statistic—a function of the outcome.) The main thing is that the more muonless events recorded, the more the result favors 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 in collaboration by 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. The question is, Does this provide evidence for the existence of neutral currents?

> For Rubbia [from Harvard] there was no question about the statistical significance of the effect . . . Rubbia emphasized that "the important question in my opinion is whether neutral currents exist or not. . . . The evidence we have is a 6-standard-deviation-effect." (P. 220)

The "important question" revolved around the question of the statistical significance of the effect. I will refer to it as the *significance question.* It is this:

> Given the assumption that the pre-Glashow-Weinberg-Salam theory of weak interactions is valid (no neutral currents), then what is the probability that HWPF would have an experiment with as many recorded muonless events as they did? (P. 220)

Three points need to be addressed: How might the probability in the significance question be interpreted? Why would one want to know it? and How might one get it?

Interpreting the Significance Question: What is being asked when one asks for the probability that the HWPF group would have an experiment with as many recorded muonless events as they did, given no neutral currents? In statistical language the question is, How (statistically) significant is the number of excess muonless events? The general concept of statistical significance will be taken up later (e.g., in chapter 5). Here I want to informally consider how it might be interpreted.

The experimental result, we said, was the recorded ratio of muonless to muonful events, namely, 54/56. The significance question, then, is, What is the probability that the HWPF group would get as many as (or more than) 54 muonless events, given the hypothesis that there are no neutral currents? The probability, notice, is not a probability of the hypothesis, it is the probability of a certain kind of experimental outcome or event. The event is that of recording *as large* a ratio of muonless to muonful events as the HWPF group did in this one experiment. It refers not only to this one experimental result, but to a set of results—54 *or more* muonless events. Wanted is the probability of the occurrence of this event given that there are no neutral currents. One way to cash out what is wanted is this: how often, in a series of experiments such as the one done by the HWPF group, would as many (or more) muonless events be expected to occur, given that there are no neutral currents?

But there is only one actual experimental result to be assessed, not a series of experiments. True, the series of experiments here is a kind of hypothetical construct. What we need to get at is why it is perceived as so useful to introduce this hypothetical construct into the data analysis.

What Is the Value of Answering the Significance Question? The quick answer is that it is an effective way of distinguishing real effects from artifacts. Were the experiment so well controlled that the only reason for failing to detect a muon is that the event is a genuine muonless one, then artifacts would not be a problem and this statistical construct would not be needed. But artifacts are a problem. From the start a good deal of attention was focused on the backgrounds that might fake neutral currents (p. 177). As is standard, one wants to assess the maximum amount of the effect for which such backgrounds are likely to be responsible and then "subtract them out" in some way. In this case, 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" (p. 217, fig. 4.40).

The problem, then, is to rule out a certain error: construing as a genuine muonless event one where the muon simply never made it to the spectrometer, and thus went undetected. To relate this problem to the significance question, let us introduce some abbreviations. If we let hypothesis H be

> H: 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.

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?

The answer is the *statistical significance level* of the result.[24]

Returning to the relevance of knowing this probability, suppose it was found to be high. That is, suppose that as many or even more muonless events would occur frequently, say more often than not, even if *H* is false (and it is simply an artifact). What is being supposed is that a result as or even more favorable to *H* than the actual HWPF result is fairly common due not to neutral currents, but to wide-angle muons escaping detection. In that case, the HWPF result clearly does *not* provide grounds to rule out wide-angle muons as the source. Were one to take such a result as grounds for *H*, and for ruling out the artifact explanation, one would be wrong more often than not. That is, the probability of erroneously finding grounds for *H* would exceed .5. This would be a very unreliable way to proceed. Therefore, a result with a high significance level is an unreliable way to affirm *H*. Hence, results are not taken to indicate *H* unless the significance level is very low.

Suppose now that the significance level of the result is very low, say .01 or .001. This means that it is extremely improbable for so many muonless events to occur, if *H* were false and the HWPF group were really only observing the result of muons escaping. Since escaping muons could practically never be responsible for so many muonless events, their occurrence in the experiment is taken as good grounds for rejecting the artifact explanation. That is because, following an argument from error, the procedure is a highly reliable probe of the artifact explanation. This was the case in the HWPF experiment, although the significance level in that case was actually considerably smaller.

This result by itself is not grounds for *H*. Other experiments addressing this and other artifacts are needed. All I am showing, just now, is the relevance of answering the significance question for ruling out an artifact. But how do you get the probability needed for this answer?

How Is the Significance Question Answered? The reasoning just described does not require a precise value of the probability. It is enough to know that it is or is not extremely low. But how does one arrive at even a ballpark figure? The answer comes from the use of various canonical statistical analyses, but to apply them (even qualitatively) requires in-

[24] Here the "null hypothesis" is that *H* is false (i.e., not-*H*).

formation about how the artifact in question could be responsible for certain experimental results. Statistical analyses are rather magical, but they do not come from thin air. They send the researcher back for domain-specific information. Let us see what the HWPF group did.

The data used in the HWPF paper are as follows (p. 220):

Visible muon events	56
No visible muon events	54
Calculated muonless events	24
Excess	30
Statistical significant deviation	5.1

The first two entries just record the HWPF result. What about the third entry, the calculated number of muonless events? This entry refers to the number calculated or expected to occur because of escaping muons. Where does that calculation come from? It comes from separate work deliberately carried out to find out how an event can wind up being recorded "muonless," not because no muon was produced (as would be the case in neutral currents), but because the muon never made it to the detection instrument.

The group from Harvard, for example, created a computer simulation to model statistically how muons could escape detection by the spectrometer by exiting at a wide angle. This is an example of what is called a "Monte Carlo" program.

> 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. (P. 217)

In short, the Monte Carlo simulation afforded a way (not the only way) of answering the significance question.

The reason probability arises in this part of the analysis is not because the hypothesis about neutral currents is a statistical one, much less because it quantifies credibility in *H* or in not-*H*. Probabilistic considerations are deliberately *introduced* into the data analysis because they offer a way to model the expected effect of the artifact (escaping muons). 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. In several places, Galison brings out what I have in mind:

> One way to recapture the lost ability to manipulate the big machines has been to simulate their behavior on a computer. In a sense the computer simulation allows the experimentalist to see, at least

> through the eye of the central processor, *what would happen* if a larger spark chamber were on the floor, if a shield were thicker, or if the multiton concrete walls were removed. (P. 265; emphasis added)

> The Monte Carlo program can do even more. It can simulate situations that *could never exist in nature*. . . . Such altered universes do work for the experimentalist. One part of the Gargamelle demonstration functioned this way: suppose the world had only charged-current neutrino interactions. How many neutral-current *candidates* would there be? Where (statistically) would they be in the chamber? (Ibid.)

Returning to the specific analysis, it was calculated that 24 muonless events would be expected in the HWPF experiment due to escaping muons. This gives the number expected to be misinterpreted as genuinely muonless.

> They wanted to know how likely it was that the observed ratio of muonless to muon-ful events (54/56) would fall within the statistical spread of the calculated ratio (24/56), due entirely to wide-angle muons. (P. 220)

They wanted to "display the probability" (as the report put it) that the difference between the number of observed and expected muonless events was merely an ordinary chance fluctuation. The difference between the ratio observed and the ratio expected (due to the artifact) is 54/56 − 24/56 = 0.536. How improbable is such a difference even if the HWPF group were experimenting on a process where the artifact explanation was true (i.e., where recorded muonless events were due to escaping muons)? This is "the significance question" again, and finally we can answer it.

What would it be like if the HWPF study actually was an experiment on a process where the artifact explanation is true? The simulation lets us model the relevant features of what it would be like: it would be like experimenting on (or sampling from) a process that generates ratios (of *m* events to *m*-less events) where the average (and the most likely) ratio is 24/56. (This corresponds to the hypothetical sequence of experiments we spoke of.) This value is just an average, however, so some experiments would yield greater ratios, others smaller ratios. Most experiments would yield ratios close to the average (24/56); the vast majority would be within two standard deviations of it. The statistical model tells us how probable different observed ratios are, given that the average ratio is 24/56.[25] In other words, the

[25] Of course, this would be correct only if the tests were at least approximately independent.

statistical model tells us what it would be like to experiment on a process where the artifact explanation is true; namely, certain outcomes (observed ratios) would occur with certain probabilities. In short, information about "what it would be like" is given by "displaying" an *experimental distribution.*

Putting an observed difference between recorded and expected ratios in standard deviation units allows one to use a chart to read off the corresponding probability. The standard deviation (generally only estimated) gives just that—a standard unit of deviation that allows the same standard scale to be used with lots of different problems (with similar error distributions). Any difference exceeding two or more standard deviation units corresponds to one that is improbably large (occurring less than 3 percent of the time).

Approximating the standard deviation of the observed ratio shows the observed difference to be 5.1 standard deviations.[26] This observed difference is so improbable as to be off the charts; so, clearly, by significance test reasoning, the observed difference indicates that the artifact explanation is untenable. It is practically impossible for so many muonless events to have been recorded, had they been due to the artifact of wide angle muons. The procedure is a reliable artifact probe.

This analysis is just one small part of a series of experimental arguments that took years to build up.[27] Each involved this kind of statistical data analysis to distinguish real effects or signals from artifacts and to rule out key errors piecemeal. They are put together to form the experimental arguments that showed the experiment could end. I would be seriously misunderstood if I were taken as suggesting that the substantive inference is settled on the basis of a single such analysis. Nothing could be further from my intent.

As Galison points out, by analyzing the HWPF data in a different manner, in effect, by posing a different question, the same data were seen to yield a different level of statistical significance—still highly significant. The error statistics approach does not mandate one best approach in each case. Its principal value is that it allows different analyses to be understood and scrutinized. Galison's excellent narration of this episode reveals a hodgepodge of different results both on the same and different data, by the same and different researchers at different

[26] The standard deviation is estimated using the recorded result and a standard statistical model. It equals $\frac{24}{56}\sqrt{\frac{1}{24}+\frac{1}{56}} = 0.105$ (Galison 1987, 220–21).

[27] The recent inference to the identification of so-called top quarks followed an analogous pattern.

times in different labs. This calls for just the kinds of tools contained in a tool kit of error statistics.

Some Contrasts with the Bayesian Model

The Bayesian model is neater, but it does not fit the actual procedure of inquiry. The Bayesian model requires the researchers to start out with their degrees of belief in neutral currents and then update them via Bayes's theorem. It also requires assessing their strength of belief in all the other hypotheses that might explain some experimental result, such as the artifact explanation of escaping muons. The researchers did not do this.

As Galison shows, different institutes at different times came up with different estimates of parameters of interest. No one is surprised by this, and, more importantly, the researchers can use these reports as the basis for criticism and further work. Imagine, in contrast, different institutes reporting their various posterior degrees of belief in *H*, neutral currents. Suppose one institute reports that the degree of belief in *H* is low. Lindley says all the information resides in an agent's posterior probability. But it is not clear what other institutes could make of this. For one thing, one would not know whether it was due to a highly discrepant result or a small prior degree of belief. A two-standard-deviation difference and a ten-standard-deviation difference indicate different things to the practitioner, but they could both very well yield an identical (low) posterior. The posterior would not have provided the information the researchers actually used to learn things such as what to do next, where the source of error is likely to lie, how to combine it with other results, or how well the data accord with the model.

Bayesians, at least officially, reject the use of significance tests and other error probability methods. (Indeed, it is hard to see how one can be a consistent Bayesian and *not* reject them. See chapter 10.) Howson and Urbach (1989), following the Bayesian fathers cited earlier, maintain that "the support enjoyed by [error statistics methods] . . . among statisticians is unwarranted" (p. 198). They declare that one of the staples of the experimenter's tool kit for assessing "goodness of fit" for a model to data (the chi-square test) "should be discarded" (p. 136)! Their criticisms, to be taken up later, stem from the fact that error statistics methods aim to perform a very different role from the one envisaged in the Bayesian model of inductive inference.

Error probabilities are not final evidential-relation measures in hypotheses. However, error probabilities of the experiment from which a claim is arrived at perform a much valued service in experiments. They provide for an objective communication of the evidence and for debate over the reasons a given claim was reached. They indicate what experiments have been performed and the process by which the estimate or result came about. They can be checked by experimenting with a different type of test. Scientists obviously find such information valuable.[28] Their value is as part of an iterative and messy series of small-scale probes using a hodgepodge of ready-to-use and easy-to-check methods. (Much like ready-to-wear [versus designer] clothes, these "off the shelf" methods do not require collecting vast resources before you can get going with them.) They will not appeal to ultra neatnicks.

By working with the data and arguments of specific cases, however, it is possible to see how the messiness of a host of piecemeal analyses gives way to rather neat strategies. The ingredients, for at least several important cases, I maintain, are already available in the works of the New Experimentalists. This is so even for Allan Franklin's work, despite his appeal to the Bayesian Way in his proposed epistemology of experiment, for his extensive examples reveal page after page of error statistics. Separate from these experimental narratives, Franklin attempts to give a Bayesian gloss to the experimental strategies he so aptly reveals. In doing so, the actual epistemological rationale of those strategies gets lost.

Scientists Are Bayesians in Disguise (and Artists Paint by Number)

Even where the use of error-statistical methods is of indisputable value, the ardent Bayesian still withholds credit from them. What the Bayesian would have us believe is that the methods used are really disguised attempts to apply Bayes's theorem, and the Bayesian will happily show you the priors that would give the same result. We may grant that experimental inferences, once complete, may be reconstructed so as to be seen as applications of Bayesian methods—even though that would be stretching it in many cases. My point is that the inferences actually made are applications of standard non-Bayesian methods. That an after-the-fact Bayesian reconstruction is possible provides no reason to think that if the researchers had started out only with Bayesian tools they would have reached the result they did. The point may be made with an analogy. Imagine the following conversation:

Paint-by-number artist to Leonardo Da Vinci: I can show that the *Mona Lisa* may be seen as the result of following a certain paint-by-number kit that

[28] This is what endears these methods to practitioners. See, for example, Lucien LeCam 1977 and B. Efron 1986.

I can devise. Whether you know it or not you are really a painter by number.

Da Vinci: But you devised your paint-by-number *Mona Lisa* only by starting with my painting, and I assure you I did not create it by means of a paint-by-number algorithm. Your ability to do this in no way shows that the paint-by-number method is a good way to produce new art. If I were required to have a paint-by-number algorithm before beginning to paint, I would not have arrived at my beautiful *Mona Lisa.*