from science to science". And he holds that one can make this concession too without surrendering to relativism. Allegedly (p. 248), those who defend 'creation science' are really claiming that the methods employed in that field are similar in certain respects to those involved in physics. But that similarity claim can be assessed by inspection of the two fields: "No universal account of science is necessary" (p. 248). But what if the claim were, as it sometimes indeed is, that the methods of 'creation science', although different from those of physics are 'equally valid'? That the methodological injunction that no theory can be satisfactory if it clashes with a 'literal interpretation' of *Genesis* is 'just as valid' as the methodological injunction that no theory can be satisfactory if it is massively *ad hoc* (while a non-*ad hoc* rival for the same range of phenomena exists)? It is this sort of suggestion that reveals that defenders of scientific rationality *must* assert firmly that there are general principles governing not just physics but all sensible empirically-based attempts to acquire knowledge. It is because it contradicts those general principles that so-called creation science is pseudoscience.

There is *a* thing called science. Despite its many imperfections in practice, blessed be its name! Reading Alan Chalmers' book will give students a good start towards understanding what that thing is (though they should resist taking Chapters 11 and 16 seriously).

Centre for Philosophy of Natural and Social Science,
London School of Economics,
Houghton Street,
London WC2A 2AE,
UK.

*By Deborah G. Mayo**

The more philosophers of science have turned their attention to historical episodes in science and to the complexities of actual scientific practice, the more they have come to see the inadequacies in all philosophical accounts of scientific evidence, inference, and hypothesis testing. Attempts to set out formal rules or logics relating statements of evidence and hypotheses by logical relations of confirmation, support, corroboration, and the like either fail to capture actual scientific inference or lack any normative force—or both. Nor have these problems been solved by the attempts to look away from logics of evidence to developing, instead, methodologies for large-scale changes in paradigms, research programmes, and the like. The question now is: Should philosophers of science give up on what has long been held as their primary task: to understand and justify scientific methods for assessing hypotheses on the basis of empirical data? And if they should, what job is left for the philosopher of the epistemology of science? What is this thing called philosophy of science?

Anyone seeking a clear, sophisticated and impressively concise tour of the developments that have led contemporary philosophers of science to this predicament will find Alan Chalmers' third edition of his *What is This Thing Called Science?* a treasure. After taking the reader through the twists and turns of the logics of confirmation and falsification, and the attempts to locate scientific rationality in large-scale theory change, Chalmers takes up this question directly: "Since I have denied that there is a universal account of science available to philosophers and capable of providing standards for judging science...it might be concluded that the views of philosophers of science are redundant and that only those of scientists themselves are of consequence. It might be thought, that...I have done myself out of a job. This conclusion (fortunately for me) is unwarranted ...Scientists are not particularly well equipped to engage in debates about the nature and status of science...such as are involved, for example, in the evaluation of creation science" (p. 252). But after finishing the book, the reader is left in the dark as to how Chalmers' philosopher of science could engage this task. Those who have grappled with it appeal to some kind of general criterion for what counts as a science (for example, falsifiability, testability), and Chalmers claims that (like Feyerabend) he denies there are general "standards that all sciences should live up to if they are to be worthy of the title 'science'" (p. 161).

At most, Chalmers allows, there are "commonsense" universals as exemplified by such general principles as "take argument and the available evidence seriously" (p. 171), but, as he concedes, it hardly takes a professional philosopher to formulate such bland generalities. How then does Chalmers think philosophers of science can help adjudicate "controversies about the nature and status of science"? They can do so, he proposes, by describing historical episodes from acknowledged sciences (for example, physics) in the right ways—the ways that emphasise the epistemological aspects of the episodes. Noting "the similarities and differences" between the disciplines, Chalmers claims, gives us "all that we need for a judicious appraisal" of claims that such and such [for example, creation science] is a science. But does it? How could one pinpoint the relevant features that must be exemplified by an enterprise before it can pass our test (of whether to count it as a science), if Chalmers is correct to deny such things exist? He nowhere answers this question. What leads Chalmers to deny there are overarching principles above and beyond trivial commonplace generalities is that once we demand more detail, "then those details will vary from science to science and from historical context to historical context" (p. 172). But, if this is so, then the mere fact that there are dissimilarities between the enterprise in question, call it *x*, and the particular historical episodes from physics that Chalmers' philosopher of science describes, cannot carry any weight.

Even though enterprise *x* has (or lacks) features found in Chalmers' favourite episodes from physics, they may well be features found (or lacking) in some perfectly good science, or in a different episode of physics—or so a defender of the status of *x* could rightly argue. Without an adequate account of 'good evidence' and what is required to 'take evidence seriously', I see no 'judicious' way to rule on the scientific merits of enterprise *x*. Fortunately, the assumption that leads Chalmers into this predicament—that the variety and context-dependencies of actual inferences preclude non-trivial norms—has much more to do with the fact that philosophers have not developed adequate accounts of evidence and inference than to a lack of non-trivial norms. Unfortunately, Chalmers does not consider this possibility. For the balance of this review, I will raise and address the following: the example of transgenic pollen, subjective Baysianism, the new experimentalism, error-statistical testing and expecially the matter of severe tests of hypotheses.

So, to cite an example, it has recently been reported that there is evidence that transgenic corn pollen (pollen from corn genetically altered to control certain pests) harms the larvae of Monarch butterflies. Here is a data report: "Larval survival after four days of feeding on leaves dusted with [transgenic] pollen was significantly lower than survival either on leaves dusted with untransformed pollen or on control leaves with no pollen" (*Nature 399*, May 20 1999, p. 214).

In particular, the observed difference in survival rates was improbably far from what would be expected by chance variability alone. This improbability, called the p-value (or statistical significance level), is given as 0.008. By contrast, had the observed difference been fairly likely even if the mortality rates were unaffected by the transgenic pollen, if, say, the p-value had been 0.4, then the data would not be good evidence of an increased mortality rate, even in the conditions of the laboratory. The former method is, and the latter method is not, a fairly reliable indicator of a genuine difference in mortality rates. While the availability and applicability of methods of interpreting data will vary this does not alter the properties of the methods (for example, reliable or not), which are objective, empirical ones. Were statistically significant increases in mortality rates interpreted as evidence that there is no risk posed—perhaps on the grounds of a very strong prior degree of belief that transgenic corn poses no increased risks to untargeted species—then that would be an example of not taking the evidence seriously. (I am distinguishing this inference about the laboratory Monarchs from an inference as to what risks are posed in the field—something that requires further errors to be ruled out.) I see nothing historically or contextually relative about this and similar principles. Although this is just a particular illustration of a method from standard statistical practice (and I can only be very sketchy in describing it here), it illustrates what is true for methods and strategies for obtaining and interpreting data in general.

There is plenty of evidence that Chalmers agrees, and it is much to his credit that he shows (in Chapter 12) that the subjective Bayesian emperor has no clothes. Especially when they are responding to criticism, subjective Bayesians stress the extent to which both the prior probabilities and the evidence which needs to be fed into Bayes' theorem are subjective degrees of belief about which the subjective Bayesian has nothing to say. But to what extent can what remains of their position be called a theory of scientific method (p. 192)? "A good theory of scientific method...will surely be required to give an account of the circumstances under which evidence can be regarded as adequate, and be in a position to pinpoint standards that empirical work in science should live up to" (p. 191).

Indeed. But for this criticism to have any weight, Chalmers needs to show how we can pinpoint (normative) standards—the very thing he suggests we are in no position to do. Can an appeal to error statistical methods rather than to Bayesian ones be the basis for a more adequate account of scientific method and inference? In discussing the "New Experimentalism" (Chapter 13), Chalmers occasionally hints that it might.

A theme running through the New Experimentalism chapter is that in place of the familiar logics of evidence (confirmation theories and inductive logics) we should focus on how experimental knowledge is actually arrived at and how it functions in science. Promising as much of this work has been, nothing like a general account of evidence and inference has been forthcoming. The reason the New Experimentalists have come up short, it seems to me, is that the aspects of experiment that have the most to offer in building an account of evidence and inference are still largely untapped: designing, generating, modelling and analysing experiments and data, activities that receive structure by means of standard statistical methods and arguments. The New Experimentalists (while hardly a homogeneous group), seem, by and large, to be too haunted by the ghosts of probabilistic logics of evidence to appeal to statistical methodology altogether. But the methodology they are forfeiting is crucially different from the logics of evidence. For one thing, rather than start with given evidence these methods direct themselves to the tasks of generating modelling and analysing data to obtain evidence in the first place. Second, in striking contrast to the logics of evidence or confirmation, probability arises in these statistical methods, not to measure degrees of credibility or support to hypotheses, but to characterise properties of tests and estimation procedures: how reliably a test is able to detect a given type of error, namely, the test's error probabilities or error characteristics. An account of evidence and testing based on error probabilities may be called an error-statistical account.

Although Chalmers looks favourably on the fruits of the error statistical account (for solving a variety of problems) the reader is not told that there is a well-worked-out battery of statistical techniques that serves as the fundamental basis for those methods. This is very unfortunate: there is a pressing need to bring these methods more squarely into philosophy of science—not just for their value to the philosophical tasks of evidence, but also, in 'the other direction' as it were, to help disentangle a host of philosophical conundrums faced by users of these methods (especially in sciences where the uncertainties are greatest). Given the alleged commitment of contemporary philosophers to the actual practices of science, it is especially surprising to find philosophers overlooking a standard set of inferential methods rather than trying to understand when and why scientists find them so useful. The time is ripe to remedy the situation: the conglomeration of statistical techniques (Neyman-Pearson tests and confidence intervals, Fisherian tests, non-parametric methods, data analysis and others) is the place to look for erecting an adequate philosophy of evidence and inference.

Granted, using these techniques to build a philosophy of evidence requires a good deal of work above and beyond any statistical texts. Most broadly put, the task for philosophers of science is to consider how to relate statistical hypotheses tests, and methods of data generation and modelling, to substantive scientific hypotheses and actual, messy, data. There is an overarching goal that may guide us in articulating these statistical-substantive links, the desire for severe tests, for severely learning from error. Impressively, Chalmers arrives at the key idea with non-technical ease: "[a] key idea ... is that a claim can only be said to be supported by experiment if the various ways in which the claim could be at fault have been investigated and eliminated" (p. 199). Keeping this central and informal idea in mind can avoid many perplexities that others generate when they take the formal statement of severity out of context and rush to find (alleged) "counterexamples". Nevertheless, this informality can result in some misunderstandings. In order for a claim or hypothesis to have passed a severe test, Chalmers writes, it "must be such that the claim would be unlikely to pass it if it were false" (p. 199). This is correct, but one must be careful not to leave off the requirement that for hypothesis H to pass the test with outcome e, H must "fit" e for an appropriate notion of fit. Chalmers omits the requirement, I think, because he is very sensitive to the fact that there are many cases where e severely passes H even though $P(e|H)$ is low, as he discusses in his appendix to Chapter 13. True, *H*'s passing a severe test with e does not require $P(e|H)$ to be high, but it must be higher than $P(e|\text{not-}H)$.[1] But how does one assess severity? Again, failing to allude to statistical methodology leaves the reader without an answer.

Although I do not wish to limit severity assessments to formal statistical hypotheses, the statistical framework gives crucial guidance for both formal and informal severity assessments. Most importantly, it teaches us that it is impossible to assess reliability or severity with just statements of data and hypotheses divorced from the experimental context in which they were generated. Minimally, we need to consider three main elements of experimental inquiry which we can represent as three types of models: models of primary scientific hypotheses, models of data, and models of experiment that link the others by means of test procedures. The primary question in our transgenic corn example above might be: does transgenic pollen harm Monarch butterflies in the field? It is tackled by asking a specific statistical hypothesis about a sample of larvae in a given experimental set-up: Is there a statistically significant increase in mortality among the sample fed transgenic pollen (treated) in contrast to the sample fed non-transgenic pollen (controls)? This is probed by considering, within a model of experiment, a standard null or error hypothesis, *Ho*, any observed difference in mortality rate between treated and control larva are 'due to chance'.

The data are modelled as the difference between mortality rates in treated and non-treated larvae. However, for the statistical inference to go through, we need to determine if they were generated in such a way that the treated and non-treated larvae are 'like random samples' from a data generation procedure where, at the start of the experiment, both groups have the same probability of mortality. By means of an interconnected set of inferences and checks, the statistical claims that are severely passed can teach about the primary hypothesis of interest.

Where tests are appropriately severe it is possible to learn from rejections and falsifications: one obtains real effects that will not go away, and in this way experimental knowledge grows. To violate the severity requirement is not to 'take evidence seriously', and if a given enterprise is regularly unable or unwilling to develop sufficiently controlled inquiries so as to distinguish different sources of error, real effects from artefacts, signal from noise, *etc.*, then it will be hindered or prevented from making progress in knowledge and its scientific credentials will be rightly questioned. In several places throughout this book, Chalmers endorses these ideas about learning from error and severe testing, and one would have thought he would put them to use in the task he regards as central for philosophers of science.

Perhaps the reason he does not is to be found in his last chapter. Here, Chalmers questions whether these ideas serve as the basis for a general account of scientific inference, because he thinks: 1) "the emphasis on experimental manipulation involved in the New Experimentalism renders that account largely irrelevant for an understanding of disciplines, especially in the social and historical sciences" (p. 250); 2) it is incomplete until it is augmented "with a correspondingly updated account of the role or roles of theory in the experimental sciences" (p. 251). Let me consider these in turn:

1) The first allegation is quite unwarranted, at least insofar as it is being alleged of the experimental account I recommend. From the very start I say "I understand 'experiment'. . .far more broadly than those who take it to require literal control or manipulation. Any planned inquiry in which there is a deliberate and reliable argument from error may be said to be experimental" (Mayo 1996, p. 7). The whole point of appealing to statistics, as I emphasise repeatedly, is that it enables us to model "what it would be like to control, manipulate, and change in situations where we cannot literally" do any of these things (Mayo 1996, p. 459). Nor are these empty promises: I make use of numerous examples that rest upon, not literal manipulation, but computer simulations, manipulations 'on paper', and other tools of the statistical trade for obtaining reliable data and severe tests by analogy to what literal controls afford. For example, if one can distinguish, through analysis, the factors responsible for a given effect, one is not hampered by being unable to hold each fixed. That is why it is so important for philosophers wrestling with problems of method to understand statistical methodology. By giving short shrift to the statistical component of the error statistical account of experiment, Chalmers completely overlooks its key features.

2) As to his second caveat, I happily accept Chalmers' urging to augment the error statistical account so as to relate "the life of experiment" to the "life of theory", and experimental knowledge to theory testing, but I would reject his suggestion that I should embrace the comparativist account of theory testing he recommends. "[Mayo's] argument for scientific laws and theories boils down to the claim that they have withstood severe tests better than any available competitor. The only

difference between Mayo and the Popperians is that she has a superior version of what counts as a severe test" (p. 208). For example, Chalmers claims I must implicitly be endorsing the position that it was warranted to accept the General Theory of Relativity (GTR), as a whole, on the basis of the eclipse results—until such time as an alternative gravity theory was available. But I do not endorse such a position. Granted, since a large-scale theory may, at any given time, contain hypotheses and predictions that have not been probed at all, it would seem impossible to say, about such a large-scale theory, that it had severely passed a test as a whole. But if one were to allow, as Chalmers recommends, that we nevertheless regard the large-scale theory as well tested, simply because no known competitor does better, one would forfeit the very fruits of the piecemeal account of severe testing that leads Chalmers to regard it as superior (for example, to Popper's). In particular, it would take us back to the problem of Popper's account of testing—namely being unable to say what is so good about the theory that (by historical accident) happens to be the best tested so far?

We can give guarantees about the reliability of the piecemeal experimental test, but we cannot give any guarantees about the reliability of the procedure: go from passing a hypothesis *H*, a proper subset of theory *T*, to passing all of *T*. Indeed, this is a highly unreliable method—anyway, it is, entirely unclear how one could ever assess this. By contrast, we can apply the severity idea because the condition "given *H* is false" (even within a larger theory) always means given it's false with respect to what it says about this particular effect or phenomenon. I am not denying that there may be licence to go from one severely tested claim to others; the ability to do so is a very valuable and powerful way of cross-checking and building on results. However, whether these connections are warranted is an empirical issue that has to be looked into on a case by case basis, whereas the comparativist is saying we are licensed to do this so long as theory *T* is the best tested so far.

The second important feature of the severity account that is given up by the comparativist (in Chalmers' sense) is that of stability. Suppose an experimental test is probing answers to a question: What is the value of this parameter? Then if a particular answer or hypothesis is severely passed, this assessment is not altered by the existence of a theory which gives the same answer to this question. More generally, in the error-statistical account of testing, if two rival theories, T_1 and T_2, say the same thing with respect to the effects or hypotheses that are being severely tested by experiment *E*, then T_1 and T_2 are not rivals with respect to *E*—no matter how much they may differ regarding domains or concepts not probed by *E*. Thus, a severity assessment can remain stable through changes in 'higher level' theories. By contrast, as soon as an alternative theory comes to light that does as well as *T* does on this (and other tests), the comparativist would regard *T* as no longer severely tested.

Eager as Chalmers (and other comparative-holists) are to license accepting an entire large-scale theory, they ignore what for our severe tester is the central engine for making progress, for getting ideas for fruitful things to do next, to learn more. Rather than asking, Given our evidence and theories, which theory of this domain is the best? we ask, Given our evidence and theories, what do we know about this phenomenon? Far from allowing ourselves to say the full theory GTR is well-tested, our severe tester would set about exploring just why we are not allowed to say that GTR is severely probed as a whole in all the arenas in which gravitational effects may occur. Even without having full-blown alternative theories of gravity in hand, we can ask (as they did in 1960): How could it be a mistake to regard the existing evidence as good evidence for GTR? To this end, a set of related experiments was modelled within what was called the parametrised post-Newtonian, or PPN, formalism. The PPN framework sets out a list of parameters that allow a systematic articulation of violations of, or alternatives to, what GTR says about specific effects. These alternatives, by the physicist's own admission, were set up largely as straw men with which to set firmer constraints on these parameters. Whereas it's not even clear, from the comparativist point of view, what motivation there would have been for deliberately erecting rival theories to GTR in 1960—after all, GTR was not facing anomalies—it is motivated from the point of view of getting more experimental knowledge about gravity, for this was the only way to extend the regions that could be said to have been severely probed. It was the only way to learn more about gravity.

Suitably massaged results of astronomical observations, organised into appropriate data models, supply the measured values of those parameters which could then be compared with the different values assigned to it by the diverse theories of gravity. In this way, in each particular solar system experiment the same PPN model of experiment mediated between the data and several alternative primary models, based on GTR and its rivals within the class of (metric) theories. What is most interesting and most deserving of greater attention are the strategies by which a primary theoretical parameter about the given post-Newtonian parameter (for example, the deflection of light by gravity) is probed by turning it into a claim or hypothesis about a statistical parameter (in the experimental model). Then, inferences from (statistically modelled) data to these statistical distributions were used to learn answers to the primary questions. Putting together the interval estimates, they constrain the values of the PPN parameters and thus squeeze the space of theories into smaller and smaller volumes. In this way they could rule out entire chunks of theories at a time (namely, all theories that predict the values of the parameter outside the interval estimate). By getting increasingly accurate estimates, more severe constraints are placed on how far theories can differ from GTR, in the respects probed. Although we may not have a clue what the final correct theory of the domain in question will look like, the experimental knowledge we can obtain now gives us a glimpse of what a 'correct' theory would say as regards to the question of current interest, no matter how different the full theory might otherwise be.

There are plenty of important philosophical issues in these inferential and modelling strategies that cry out for philosophical elucidation, but one thing is for sure: by turning a blind eye toward the methods and models of statistical data analysis, modelling, and inference, this thing called philosophy of science will continue to wring its hands and lament its emasculation in the face of controversies about the nature and justification of scientific knowledge.

* I gratefully acknowledge the support of an NSF Scholars Award during 1998–9. I also benefited greatly from lengthy correspondence with Alan Chalmers.

[1]. Some points about my notation: P(*e*|*H*) is not the usual "conditional probability" but rather the probability of outcome *e* under the assumption of, or according to, the assignment given in statistical hypothesis *H*. There is no prior probability assignment to *H*. "Not-*H*" is not the so-called catchall hypothesis. It is not even a disjunction of hypotheses. It is the denial of a specific hypotheses *H*, for example, if *H* asserts a parameter is less than *m*, not-*H* asserts it is greater than *m*.)

Department of Philosophy,
Virginia Polytechnic Institute and State University,
Blacksburg, VA 24061,
USA.

By J. J. C. Smart

Alan Chalmers' book is a truly excellent introduction to the philosophy of science. He writes lucidly and with a charming (but I think excessive) modesty, and he is able to make use of his early experience when he was an experimental physicist. A large part of the book is a fine critique of the ideas of Popper, Kuhn, Lakatos and Feyerabend, and the last three of them at least were heavily concerned with illuminating and testing ideas about the nature of science and its