RMM Vol. 2, 2011, 79–102
Special Topic: Statistical Science and Philosophy of Science
Edited by Deborah G. Mayo, Aris Spanos and Kent W. Staley
http://www.rmm-journal.de/

Deborah G. Mayo

Statistical Science and Philosophy of Science: Where Do/Should They Meet in 2011 (and Beyond)?

1. Introduction

Debates over the philosophical foundations of statistics have a long and fascinating history; the decline of a lively exchange between philosophers of science and statisticians is relatively recent. Is there something special about 2011 (and beyond) that calls for renewed engagement in these fields? I say yes. There are some surprising, pressing, and intriguing new philosophical twists on the long-running controversies that cry out for philosophical analysis, and I hope to galvanize my co-contributors as well as the reader to take up the general cause.

It is ironic that statistical science and philosophy of science—so ahead of their time in combining the work of philosophers and practicing scientists[1]—should now find such dialogues rare, especially at a time when philosophy of science has come to see itself as striving to be immersed in, and relevant to, scientific practice. I will have little to say about why this has occurred, although I do not doubt there is a good story there, with strands colored by philosophy, sociology, economics, and trends in other fields. I want instead to take some steps toward answering our question: Where and why should we meet from this point forward?I begin with some core themes, concepts, and questions, and with why we have collected statisticians, econometricians, and philosophers of science in a room for two days in June at the London School of Economics (it was air-conditioned!) to further the dialogue.

1.1 Meeting on a Two-Way Street

Despite the challenges and changes in traditional philosophy of science, at least one of its central jobs is or ought to be to clarify and help resolve the conceptual, logical, and methodological discomforts of scientists, especially in a field like statistics, which deals with issues of scientific knowledge, evidence, and inference. So philosophy of science should be relevant to foundational debates in statistics. At the same time, philosophers are interested in solving long-standing problems about evidence and inference, and ideas from probability and statistics have long been appealed to for that purpose. So advances in statistical science should be relevant to philosophy of science. Now, few philosophers of science doubt that science is successful and that it makes progress. A core philosophical problem, to put it in untraditional terms, is how to justify this *lack* of skepticism. Any adequate explanation of the success of science would have to square with the fact of limited data, with unobserved and unobservable phenomena, with theories underdetermined by data, and with all of the slings and arrows of the threat of error.

As such, we might put the central question of relevance to both philosophy of science and statistics as: How do we learn about the world *despite limited data and threats of error*?

[1] See for example the proceedings in Godambe and Sprott 1971; Harper and Hooker 1976.

1.2 Inductive Inference as 'Evidence Transcending'

The risk of error enters because we want to find things out—reach claims or take action—based on limited information. As with any inductive argument, we want to move beyond the data to claims that are 'evidence transcending'. The premises can be true while the conclusion inferred may be false—without a logical contradiction. Conceiving of inductive inference, very generally, as 'evidence-transcending' or 'ampliative' reasoning frees us to talk about induction without presupposing certain special forms this can take. Notably, while mathematical probability arises in inductive inference, there are two rival positions as to its role:

- to quantify the degree of confidence, belief, or support to assign to a hypothesis or claim given data x; and
- to quantify how reliably probed, well-tested, or corroborated a claim is given data x.

This contrast is at the heart of a philosophical scrutiny of statistical accounts. In the first, an inference to H might be merited to the extent that H is highly probable; in the second, to the extent that H is highly probed; alternatively, the goal might be in terms of comparatively probable, or comparatively severely probed. However, there are different ways to cash out the entry of probability under both headings; so I deliberately leave them in a rough form for now.

1.3 Relevance for Statistical Science Practice

I would never be so bold as to suggest that a lack of clarity about philosophical foundations in any way hampers progress in statistical practice. Only in certain moments do practitioners need a philosophical or self-reflective standpoint. Yet those moments, I maintain, are increasingly common.

Even though statistical science (as with other sciences) generally goes about its business without attending to its own foundations, implicit in every statistical methodology are core ideas that direct its principles, methods, and interpretations. I will call this its *statistical philosophy*. Yet the same statistical method may and usually does admit of more than one statistical philosophy. When faced with new types of problems or puzzling cases, or when disagreement between accounts arises, there is a need to scrutinize the underlying statistical philosophies. Too often the associated statistical philosophies remain hidden in such foundational debates, in the very place we most need to see them revealed. But, then, we need to elucidate precisely what it means to scrutinize a statistical method philosophically (*section 4.1*).

1.4 Joyful Eclecticism or a Mixture of Conflicting Methods?

From one perspective, we may see contemporary statistics as a place of happy eclecticism: the wealth of computational ability allows for the application of countless methods with little hand-wringing about foundations. Contemporary practitioners may work blissfully free of the old frequentist-Bayesian controversies; younger statisticians, even when waxing philosophical, seem only distantly aware of them. Doesn't this show that we may have reached 'the end of statistical foundations'?

My take is just the opposite. Only through philosophical scrutiny can we understand the wealth of formal machinery and, most especially, critically appraise its. Some statisticians suggest that throwing different and competing methods at a problem is all to the good, that it increases the chances that at least one will be right. This may be so, but one needs to understand how to interpret competing answers and relate them to one another, which takes us back to philosophical underpinnings.

1.5 Even Shallow Drilling Reveals Issues of Statistical Philosophy

One need not drill too far below the surface of many contemporary discussions of statistical method in order to discern a deep (and also deeply interesting) lack of clarity, if not unease, at the foundational level. Today's debates clearly differ from the Bayesian-frequentist debates of old. In fact, some of those same discussants of statistical philosophy, who only a decade ago were arguing for the 'irreconciliability' of frequentist p-values and (Bayesian) measures of evidence, are now calling for ways to 'unify' or 'reconcile' frequentist and Bayesian accounts, often in the form of one or another 'nonsubjective' or 'default' Bayesian paradigms. These attempts—the debates they've triggered and our reactions to them—give us a place to begin to discuss where, on what common ground, statistical science and philosophy of science might meet.

The reasons some statisticians give for contemporary frequentist-Bayesian unifications are both philosophical and pragmatic. For one thing, despite the growing use of computerized programs that readily enable Bayesian analyses, frequentist methods have not disappeared as they were supposed to, and ensuring low error probabilities remains a desideratum scientists are unwilling to forgo. For another thing, there is the concern that methodological conflicts may be bad for the profession:

> "We [statisticians] are not blameless [...] we have not made a concerted professional effort to provide the scientific world with a unified testing methodology [...] and so are tacit accomplices in the unfortunate statistical situation." (Berger 2003, 4)

Some Bayesians claim that frequentist methods cannot deal with the complex multiparameter situations of current practice, but this is belied by leading statistical modelers (see contribution by Aris Spanos). Any lingering doubts about frequentist methods being able to handle large numbers of variables are removed by David Hendry's econometric methodology, which intertwines model discovery with an iterative series of model validation-tests, all while controlling error probabilities (see the contribution by David Hendry).

1.6 "An Important Task of Our Time"

The advantage of a frequentist-Bayesian unification, many claim, is to ensure that answers are conditional on the data actually obtained while at the same time respecting the frequentist notion that the methodology must ensure success in repeated usage by scientists (Berger 2006, 388).

However, these twin goals turn out to conflict with one another! Thus the increased use of nonsubjective Bayesianism in general, and the attempts at 'reconciliation' in particular, have, at least implicitly, put foundational issues back on the map, despite not always being noticed. Nonsubjective Bayesian methods permit violations of fundamental principles long held as integral to what subjective (or personalistic) Bayesians consider the 'Bayesian standpoint' (e.g., Lindley 1997). With good reason, leading subjective Bayesian statisticians are at the forefront in confronting their fallen brethren. It is as if some of the generals from the earlier (Bayesian-frequentist) statistics battles were wondering just who (if anyone) had won the statistics wars. Take Jay Kadane (2008, 457; emphasis added):

> "The growth in use and popularity of Bayesian methods has stunned many of us who were involved in exploring their implications decades ago. The result [. . .] is that there are users of these methods who do not understand *the philosophical basis of the methods they are using*, and hence may misinterpret or badly use the results [. . .]. No doubt helping people to use Bayesian methods more appropriately *is an important task of our time*."

I quite agree with Kadane as to the importance of the task. In addressing it, however, we must ask: Can contemporary statistical practitioners be 'helped' to use Bayesian methods in the manner deemed appropriate by the personalist founders? Is there just one philosophical basis for a given set of methods?

Clearly not. Among frequentist founders, for instance, R. A. Fisher is readily acknowledged to have embraced a philosophical foundation different from those embodied by Neyman and Pearson. Even within those schools there are competing evidential vs. behavioristic interpretations and foundations. The differences between contemporary default Bayesians and subjective Bayesians, many think, are even more dramatic than the differences between Fisherian and Neyman-Pearsonian frequentists (see Stephen Senn's contribution). Kadane's 'important task' is indeed important, and it is *philosophical*. Arguing for one rather than another way to use and interpret a given formal methodology is a crucial task for contemporary philosophy of statistics. For it is in these arguments that a statistical science-philosophy of science meeting ground, of relevance to current practice, will emerge.

1.7 The Philosophical Doctor Is In

As thorny as these philosophical problems are, we can get a handle on them by looking to a handful of questions:

- What are the roles of probability in inductive/statistical inference in science?
- What are the goals/functions of inductive/statistical inference in relation to scientific inquiry?

These queries will guide us as we reexamine the philosophical basis of the mathematical methods of statistics, both old and new. They require that we ask:

- What can various methods be used for?

The answer is distinct from what a method's founders may have had in mind, and from textbook accounts. It demands that we stand 'one level removed' from common interpretations and applications of methods. For example, Bayesian methods may be adequate for updating prior degrees of belief in an exhaustive set of hypotheses, but many deny that this is the only or even the best use of these methods. Likewise, standard frequentist methods, e.g., hypotheses tests and confidence interval estimation procedures, may be adequate for the goal of ensuring low long-run frequencies of erroneous inferences (or decisions), but they may be used for rather different goals in the contexts of the scientific inquiry which is my focus.

I do not want to rehash the 'statistics wars' that have raged in every decade from the 1960s to the present, even though the so-called 'significance test controversy' is still hotly debated among practitioners (in psychology, epidemiology, ecology, economics), and even though it can sometimes seem that each generation is fighting these wars anew—with journalistic reforms, and with task forces set up to stem reflexive, recipe like uses of statistics that have long been deplored. I have discussed these debates at length elsewhere, and although this discussion will have implications for resolving them that is not where I propose to begin today.[2] If we are to make progress in resolving these decades-old controversies, which still shake the foundations of statistics, as well as tackle new ones, we need to dig (or drill?) not shallowly but deeply, a task that requires both statistical and philosophical acumen. The drilling analogy seems especially apt given the obsession (in the U.S.) with the oil spill in the Gulf of Mexico during the 2010 summer of our initial forum, and I retain it here.

The job of the philosopher is to clarify but also to provoke reflection and scrutiny precisely in those areas that go unchallenged in our ordinary lives and practices. My remarks may well be provocative to all existing sides of the debate about the roles of probability and statistics in scientific inquiry and learning.

2. Induction and Error

2.1 Probability and Induction

Whether probability purports to be used to quantify degrees of belief/support, or to capture degrees of well-testedness/corroboration, or the like, we do not have

[2] See for example Mayo 1985, 1992, 1996; Mayo and Cox 2010; Mayo and Spanos 2006, 2010, 2011.

an inductive inference until we *detach* some claim or assertion (be it probabilistic or other). The following conditional, for example, would not be considered an inductive inference:

> If a weighing experiment is adequately modeled as independent and identically distributed (iid) random variables from a normal distribution with mean μ, standard deviation σ, then the probability of the 95% confidence interval estimation procedure containing the true value of μ is .95.

Even adding a prior-probability distribution to this same conditional, and deducing a posterior probability for parameter(s) μ (and/or σ), is not yet to make an inductive inference, as I am using the term. Once it is granted (ideally by adequate checking) that the antecedent assumptions of the model (and in the latter case, the prior) hold, various inductive inferences are possible. It is true that 'inference' can refer to the entire argument or to the particular conclusion, but that is not the point. The conclusion inferred, to be genuinely inductive (or ampliative), must take the leap of going beyond the premises.

An inductive inference from a standard frequentist method might take the form:

(i) The data indicate that my weight is 130 pounds (generally with a specific approximation interval given).

Or the inference might just be the detached claim:

(ii) My weight is less than 130 pounds,

accompanied by the reliability characteristics of the estimation procedure.

A Bayesian inference might take the form:

(iii) The posterior probability that my weight is less than 130 pounds is .95.

Many other variations of both frequentist and Bayesian inferences are possible.

Both accounts require background information to arrive at the model for the phenomenon, to specify the data generation technique, and to check the adequacy of a statistical model for data x. These tasks demand their own inferences.

2.2 Statistical Science: Learning Despite Error

We deliberately used 'statistical science' in our forum title because it may be understood broadly to include the full gamut of statistical methods, from experimental design, generation, analysis, and modeling of data to using statistical inference to answer scientific questions. (Even more broadly, we might include a variety of formal but nonprobabilistic methods in computer science and engineering, as well as machine learning.) Since statistical science directs itself to achieving these tasks in the face of limited information, uncertainty, and error, it stands to reason that its methods would be relevant to the general philosophical one (one of the arrows on the two-way street).

Statistical methods, as I see them, provide techniques for modeling, checking and avoiding, and learning from these mistakes. This conception of statistics is sufficiently general to embrace any of the philosophies of statistics now on offer, even though each requires its own interpretation (to which we will return). It does not readily lend itself to a single overarching 'logic' of the sort to which philosophers of science sometimes look. The difference between these empirical and highly context-dependent uses of statistical methods, and the philosophical pastime of erecting overarching logics to relate evidence statements and hypotheses, reveals an obstacle to finding a meeting ground for philosophy of science and statistical science. Only by removing this obstacle can statistical ideas be used to solve problems philosophers care about, which gives us a shot at obtaining an account of ampliative inference relevant to actual scientific learning.

2.3 Twin Goals: Reliability and Informativeness

While philosophers tend to draw skeptical lessons from the fact that error is always possible, statistical practitioners focus on specific threats to the validity of their inferences and claims of evidence. Philosophers of science can learn from this: if we want to understand how we manage to be so successful despite the threat of error, we should look not at the worst cases but at where and how humans learn despite error. The fundamental role of statistical concepts and methods, as I see it, is to provide a growing machinery to capture and cope with some canonical types of errors that arise across a wide range of areas.

On the one hand, we want a method that recognizes the error-proneness of inductive learning; on the other, we do not want the error-control to be so extreme that little of informative significance is learned. Another way to put this is that we want both *reliability (of tests) and (informativeness) of claims inferred*. Focusing on the character of error probing, discriminating, amplifying, and learning from error seems a promising way to locate essential features of inductive learning.

I do not mean formal statistical errors, but general mistakes in inference, such as erroneously inferring a genuine (as opposed to a spurious) effect, mistakes about parameters (whether in a theory or a statistical model), mistakes about causal processes or mechanisms, and mistakes about the adequacy of a model—both for arriving at a statistical inference, and, separately, for learning about some phenomenon of interest.

2.4 Frequentist Error Statistics

Frequentist statistics employs the frequentist notion of probability, but to say this is scarcely to capture its essential ingredients. The key ingredient, just from the formal statistical perspective, is the use of probabilities to quantify the error rates in applying a (test or estimation) procedure. For instance, a significance test T appeals to probability to assess the proportion of cases in which a null hypothesis H_0 would be rejected in a hypothetical series of repeated uses of test T, when in fact H_0 is true. This is an error probability. Note that an error probability is associated with *a method* for inference or testing.

Imagine I weighed in at 130 pounds before my trip to London and I wish to investigate if there has been any weight gain upon returning, using a number of scales with known precisions. My weight is an unknown fixed parameter (at this moment), as would be any weight increase δ. A typical null hypothesis is:

$H_0 : \delta = 0$

(or the inference may specify an upper bound to the increased weight).

In general, there is a test procedure T that leads from data on measurements, x, to hypotheses about the data generating procedure—here, my weight. T's reliability refers to notions such as: the probability test T erroneously outputs 'x indicates H_0' (no increase from 130 pounds). Here is an example of a reliable test for this case: Infer that no more than one pound has been gained only when none of the three different scales of known precision detects an increase, even though they readily discern the addition of a one-ounce potato.

A Bayesian analysis would consider a prior distribution on the unknown fixed weight. But given that the problem stipulates a fixed weight, what can the prior here be interpreted as? Bayesians might construe the prior as representing a degree of prior belief in different values I might weigh, or they might use a 'default' prior distribution. This leads to a posterior probability in H_0 that I have not gained weight. The error statistician and the Bayesian (of either stripe) are asking distinct questions. C. S. Peirce, writing in the late nineteenth century, captures the error-statistical spirit:

> "The theory here proposed does not assign any probability to the inductive or hypothetic conclusion, in the sense of undertaking to say how frequently *that conclusion* would be found true. It does not propose to look through all the possible universes, and say in what proportion of them a certain uniformity occurs; such a proceeding, were it possible, would be quite idle. The theory here presented only says how frequently, in this universe, the special form of induction or hypothesis would lead us right. The probability given by this theory is in every way different—in meaning, numerical value, and form—from that of those who would apply to ampliative inference the doctrine of inverse chances." (Peirce 1931–1935, vol. 2, para. 748)

However, the Bayesian procedure might also be construed as a general rule, just like test T, and Peirce's question might be: How frequently, in this universe, would the method lead us right?—an error-statistical query. One might ask, for example: What is the probability of a high posterior in H_0 even if it is false? This might be construed as placing an error-statistical analysis upon a Bayesian method, and some Bayesian-frequentist reconciliations take this form. Without great care as to what is varying (the random variable? Or also the parameter?), the result can differ greatly from a genuine error-statistical assessment.

2.5 Error-Statistical Methods as Tools for Severe Testing

It is ironic that gestures toward reconciling frequentist and Bayesian methods make a point of showing that recommended techniques have good success rates in repeated usage, given that the central criticism traditionally leveled at frequentist methods questions the relevance of low long-run error rates to particular inferences. The latter appeals to a 'behavioristic goal'—one will not often 'act' erroneously regarding a phenomenon in the long run—whereas we want an 'evidential' construal for the case at hand. I agree. Long-run reliability is a necessary but insufficient use of tests, and properly interpreted error-statistical tests may be used to control and scrutinize how well or *severely* tested a given hypothesis is with specific data x. (It is the unificationist promoting long-run performance who owes us a rationale!)

Consider my example of inferring an upper bound for weight gain using well-calibrated scales. While it is true that the method is reliable—that by following such a procedure in the long run one would rarely report weight gains erroneously—that is not the rationale we demand for the particular inference. Rather, the justification is that were I to have gained more than d pounds, the test would have, with high probability, revealed this in one of my checks. The claim that x is evidence that μ is less than μ', we might say, has passed a stringent or severe test. Likewise, a nonstatistically significant difference x is poor evidence for $\mu < \mu'$ if such an insignificant result would occur with high probability, even if μ were as great as μ'. In that case our assertion passes with low severity. This reflects what I consider a *minimal principle of evidence*.

Use of frequentist methods for this kind of evidential appraisal may be called a *severe testing account based on error statistics*. Although frequentist methods do not (usually) directly supply a severity assessment, they may be used for this aim, and I take that as their philosophical justification. The severity concept (any number of analogous terms might be used) supplies the formal frequentist methods with a statistical philosophy. It avoids the classic criticisms of frequentist methods while enjoying a sound foundation: it lets us determine what we have and have not learned. Its advantages with respect to the task of grounding the use of statistical models is a distinct topic which I leave to others (see the contribution by Aris Spanos). On the philosophical side, the severity interpretation of frequentist statistics might enable the right-headed element in Popper to be fruitfully implemented by current day 'critical rationalists' (see Max Albert's contribution).

3. A Platform on Which to Meet

Here then is a place to look to meet directly with a host of foundational problems current in statistics: the discussions of 'unifications' or 'reconciliations' between Bayesian and frequentist methods. Certainly it was the current work on reconciliations that opened my eyes to this latest round in the philosophy of statistics battles. There are at least two kinds of 'meetings' represented in these purported unifications: first, between frequentist and Bayesian methods, but also, between statistical methodology and epistemology of science. It is precisely the tensions to which both kinds of meeting grounds give rise that reveal where current foundational issues come up against basic philosophical assumptions (about the role of probability in inductive learning, and the role of formal statistical tools).

3.1 Frequentist-Bayesian Unifications

Granting that "agreement on statistical philosophy is not on the immediate horizon", Jim Berger will "focus less on what is correct philosophically than on 'what is correct methodologically'" (Berger 2003). His allusion to philosophical agreement seeming to be far off suggests that professional philosophers are at least wrestling with the issue. By and large, they have not been part of the contemporary debate. I hope to forge a shift in this status quo. Since a successful unification must be thought to satisfy the fundamental goals or the minimal requirements of frequentist and Bayesian accounts, looking at attempted unifications is very revealing as to presuppositions about those goals.

In one key paper, Berger (2003) purports to produce a piece de resistance: an account of testing to which Jeffreys, Fisher, and Neyman could have agreed. But if he really can produce the low long-run errors of methods that frequentists demand, then wouldn't the entire method fall under the frequentist (error-statistical) umbrella? In fact it turns out that he is using the notion of an 'error probability' in a different manner, i.e., as a posterior probability assignment to a parameter, even though the parameter is regarded as fixed.

So what shall we say to Berger's suggestion about agreeing on methodology without philosophy? If there is an agreement on numbers despite different interpretations and different intended questions being asked of data, it cannot lead to the sound professional concordance he seeks. So his task implicitly calls for foundational work.

3.2 Diffident Bayesianism

Contemporary work on Bayesian-frequentist unifications offers the frequentist error statistician a clearer and less contentious (re)entry into statistical foundations than when Bayesian 'personalists' reigned (e.g., Lindley, Savage). Confronted with the position that "arguments for this personalistic theory were so persuasive that anything to any extent inconsistent with that theory should be discarded" (Cox 2006, 196), frequentists might have seen themselves in a kind of exile when it came to foundations, even those who had been active in the dialogues of an earlier period. Sometime around the late 1990s there were signs that this was changing. Once again I will resist trying to explain why this occurred, but *that* it occurred is of central importance to statistical philosophy.

Unlike their subjectivist predecessors, the Bayesian statisticians leading the unifications favor the use of what we may call 'nonsubjective' Bayesian priors if only to avoid letting scientists' subjective beliefs overshadow the information provided by data. Here, prior probability distributions arise from a variety of formal considerations. (These nonsubjective Bayesian paradigms have their own history in statistics and philosophy, notably, in the work of Jeffreys and Carnap, respectively.)

With the early attempts, the dream of priors that leave inference pure and unadulterated could still be entertained; nowadays it is conceded (at least by statisticians) that “non informative priors do not exist” (Bernardo 1997). The old dream has been replaced by finding conventional or ‘default’ choices of prior distributions for parameters of statistical models that reflect a lack of subjective information. The impressive technical complexities notwithstanding, the result has been a multiplicity of incompatible ways to go about this, none obviously superior (Bernardo 2010).[3] So the desired ‘agreement on numbers’ has yet to materialize even within the nonsubjective Bayesian family; one may pick one technique, be it Bernardo’s or Berger’s or another’s, but an interpretation and foundation is still needed (see Jan Sprenger’s contribution).

3.3 A Plethora of Foundational Problems

By finding nonsubjective priors we can (at times) get posteriors that match error probabilities. In so doing, some claim, we both recover current (frequentist) statistical practice while giving it the philosophical foundation it lacks. The trouble is that neither holds up: the error probabilities that match the posterior may no longer supply either the frequentist error probabilities or the celebrated philosophical foundations. A word on each:

In some cases the nonsubjective posteriors may have good error-statistical properties of the proper frequentist sort, at least in the asymptotic long run. But then another concern arises: If the default Bayesian has merely given us technical tricks to achieve frequentist goals, as some suspect, then why consider them Bayesian (Cox 2006)? Wasserman (2008, 464) puts it bluntly: If the Bayes’ estimator has good frequency-error probabilities, then we might as well use the frequentist method. If it has bad frequency behavior then we shouldn’t use it. (The situation is even more problematic for those of us who insist on a relevant severity warrant.)

Subjective Bayesians are not much happier with the unifications. They focus too much on technique at the expense of the ‘Bayesian standpoint’ (i.e., updating degrees of belief, says Dennis Lindley (1997), commenting on Bernardo). Whereas in the subjective Bayesian standpoint, the fundamental role for the prior was formally to incorporate into inductive inference an agent’s degree of belief, apart from the data and statistical model, the nonsubjective priors are model-dependent, and are not even intended to represent beliefs. (They are often not even probabilities.) The recommended conventional priors lead to Bayesian incoherence, thwarting what had long been taken as the heart of Bayesian foundations. Several Bayesians complain that the cottage industry that has grown

[3] Even in simple problems, recommended Bayesian procedures differ. See the definitive review by Kass and Wasserman (1996).

up for finding default priors is taking practitioners away from more important work.

3.4 Bayesian Family Feuds

A forum in *Bayesian Analysis* (vol. 1, no. 3, 2006) exemplifies the kind of philosophical family feuding that is common in current practice, with or without non-Bayesian frequentist input (usually without). The representatives are Jim Berger and Michael Goldstein, representing default Bayesianism and subjective Bayesian practice, respectively. Remarkably, both lead papers (and others in the discussion) show the disintegration of traditional Bayesian foundations. Jim Berger's position is not uncommon: even if, in his heart of hearts, he believes that Bayesian updating provides authentic philosophical reasoning in contexts of updating subjective degrees of belief, Bayesians should, in practice, adopt some standard default priors.

"The (arguably correct) view that science should embrace subjective statistics falls on deaf ears; they come to statistics in large part because they wish it to provide objective validation of their science." (Berger 2006, 388) Subjective elicitation is not only unreliable, he feels, it detracts from the more serious problem of model specification. Further, the use of default priors combats what he terms "pseudo-Bayesian" subjectivism, wherein prior probabilities with poor performance characteristics are adopted under the banner of subjectivity.

Despite his role as defender of subjective Bayesianism, Goldstein says he "cannot remember ever seeing a non-trivial Bayesian analysis which actually proceeded according to the usual Bayes formalism". Like Berger, I find it interesting to note that he "is not making a ringing endorsement of what is perceived as standard subjective Bayesian analysis". This seems increasingly common, even when it comes to advocating the use of Bayesian updating itself: "There is no stronger reason why there should be a rule for going from prior to posterior beliefs than that there should be such a rule for constructing prior beliefs in the first place." (Goldstein 2006, 414) The need to avoid 'betting incoherency' seems to have gone by the wayside as a kind of justification, as opposed to a tautologous result for contexts where all the 'givens' are granted. The status among philosophers of probability is less clear. (Howson appears to reject appeals to Dutch Books beginning in 1997.)

While Bayesianism is appealed to for philosophical foundations, in practice it is toward reference or default Bayesian priors that many look (Kass and Wasserman 1996); so any philosophical problems it faces are relevant to a large part of current Bayesian practice, which in turn is relevant to Bayesian philosophy of science.

3.5 Disinterring Frequentist Roots?

There are plenty of practitioners wearing Bayesian hats who are not members of (or even reject) the unificationist movement. Here, too, however, there seems to be widespread disagreement about the recommended 'Bayesian' solutions in a great variety of domains. The methods advocated throw together likelihoods, priors (of all stripes), sampling distributions, conditioning, significance tests, confidence intervals, subjective and default priors, linear models, and everything else in the statistical kitchen sink.

Having abandoned the traditional foundational justifications, these Bayesians tend to defend their methods by pointing to their 'usefulness'. The question of what if any general principles, reasoning strategies, or underlying rationales are actually responsible for the results they value is left glaringly open. We cannot credit a method for a useful result without being clear that it is because of the method.

The last decade or more has also given rise to many new problem areas that call for novel methods (e.g., machine learning). Do they call for new foundations? Or, can existing foundations be relevant here too? (See Larry Wasserman's contribution). A lack of clarity on the foundations of existing methods tends to leave these new domains in foundational limbo. Some discussions reveal widespread unclarity about the nature of frequentist statistics. Bayesian critics agree on one thing: frequentist methods license a handful of 'hilarious' examples, often described just before turning to the preferred Bayesian approach (Ghosh et al. 2006).

Some statisticians describe themselves as Bayesian while at the same time advocating Fisherian statistical significance tests, and some even suggest that "the idea of Bayesian inference as inductive, culminating in the computation of the posterior probability of scientific hypotheses, has had malign effects on statistical practice" (Gelman and Shalizi 2010) (see the contribution by Andrew Gelman). Philosophers of science are legitimately called upon to sort things out. It seems altogether possible that elements of current practice are implicitly disinterring frequentist roots, even as these roots are unaccompanied by a clear recognition of their statistical philosophy.

3.6 Classic Criticisms of 'Classic' Frequentist Methods

In declaring that the philosophical doctor is in (*section 1.7*), I identified two key areas around which to organize foundational issues: the roles of probability in induction, and the nature and goals of statistical inference in science or learning. Two implicit assumptions underlie the criticisms of frequentist accounts:

First, there is the supposition that an adequate account must provide hypotheses with degrees of probability, an assumption often called *probabilism*.

Second, there is the assumption that the sole role of error-statistical methods is to appraise techniques according to their long-run error rates (however defined). This assumption may be dubbed the *radical behavioristic interpretation*.

Criticism then follows easily: Error probabilities do not give posterior probabilities to hypotheses, and methods that satisfy low long-run error probability requirements may be counterintuitive.

It will be evident that I reject both presuppositions that underlie the criticisms.

3.6.1 Probabilism

There are really only two or three variations on the ensuing criticisms. The first charge is based on the assumption that probability must arise to assess posterior probabilities. Were the results actually adequate for quantifying something like rational belief that would be one thing, but by and large, they are not. Still, it is a basic assumption that many apparently feel is not in need of justification.

Standard Bayesian textbooks make obligatory claims based on analogies with games of chance: Since probability is used to quantify how strongly an uncertain event is 'expected' to occur in the context of a probabilistically modeled game of chance, probability should also be used to quantify the evidential warrant for a hypothesis H, even where they too regard H as correct or incorrect (about this one universe). It seems to me that there is confusion between 'expecting an event' to occur and expecting a hypothesis H to be true. Even if scientists were in the business of betting on the truth of hypotheses, there is an entirely different role for statistics in ascertaining what has been learned about a given phenomenon. This is the role, I argue, for controlling and assessing how precisely or severely given hypotheses have (and have not) passed tests with data.

In some cases, of course, a parameter has a legitimate prior probability distribution. Even then, however, it is not clear that one ought to employ it for the inference at hand. Moreover, some deny that such a computation should even count as performing a Bayesian analysis, as opposed to simply applying conditional probability (Fraser forthcoming).

Trivial Intervals. One way the probabilist assumption leads to classic criticisms is by assuming that error probabilities are intended to supply post-data degrees of belief in hypotheses. So, if, for example, the result of applying a 95-percent confidence interval estimation procedure happens to be known as a true estimate, then this demonstrates 'unsoundness' of frequentist methods.

Now frequentists have been pointing out for over half a century that a confidence level is not a posterior probability assignable to a resulting estimate (which is correct or incorrect); nevertheless, they too tend to accept the criticism, or at least seem uncomfortable with these examples. An exception is David Cox, who has no problem allowing that in some cases none of the parameter values can be ruled out with any stringency. The severity construal concurs.

3.6.2 Radical Behaviorism

The second assumption gives rise to what are regarded as the strongest grounds for preferring some variety of the Bayesian to the frequentist error-statistical method: namely, that the Bayesian avoids the counterintuitive and paradoxical results that the frequentist (supposedly) licenses. In the spirit of Jim Berger, I will admit flat out that the frequentists are not blameless. With few exceptions (e.g., Cox), rarely have they mounted a strong enough response to these foundational problems and puzzles, even where critics take them as sufficient grounds for rejecting the frequentist approach altogether.

To give the most generous reading: the classic paradoxes are easy to make out if one assumes the most radical type of behaviorism, beyond anything that even Neyman endorsed in his most behavioristic moments. This assumes that the frequentist error-statistical requirement is satisfied so long as on average the method has good long-run error probabilities. So even if one scale is terrible and the others highly reliable, the frequentist, it is imagined, is happy to average them together in reporting on the warrant for a weighing, even once the scale used is known. (So using my broken scale is not too bad if I can claim that most of the time I use a reliable scale.) But why suppose the frequentist statistician is stuck advocating such counterintuitive applications? Certainly reporting the average will incorrectly report how well (e.g., how severely) the hypothesis has actually passed the test with the experiment producing x.

So, having embarked on our meeting ground, we are led to examine very carefully these old chestnuts and 'hysterical' examples laid at the frequentist, error-statistical door. Since the same counterexamples are given by Bayesians who view themselves as neither subjectivists nor default Bayesians nor unificationists, the analysis is widely applicable to the foundational portions of all Bayesian textbooks.

4. How to Scrutinize Statistical Methods Philosophically

Both statisticians and philosophers of science have an interest in scrutinizing the philosophical basis of a statistical or other inductive method (even if they go about it in different ways). Here, one asks how to interpret and justify the method in the light of its intended goals. Given our focus on science and learning, this involves epistemological goals—goals of learning or knowledge or, as I prefer, simply finding things out. In scrutinizing a statistical account at a philosophical or foundational level, we (do or should) ask: *Does it provide an adequate characterization of scientific reasoning, evidence, inference, testing?*

4.1 Criteria for the Philosophical Scrutiny of Methods

I propose some criteria for answering this question. A first pair of requirements for an adequate methodology are:

1. *It should be ascertainable (it must be able to be applied); and*
2. *It should be adequate and relevant to the tasks required of the inference tools.*

These two criteria are interrelated: If it is assumed that an adequate account must supply posterior probabilities to hypotheses, then a frequentist account that assigns only probabilities to events will fall down on the adequacy criterion. However, for posterior probabilities to be ascertainable, it is necessary to give prior probability assignments to all possible hypotheses that might give rise to the data so as to apply Bayes's theorem. The question arises as to how to understand these priors so as to be both ascertainable and relevant to scientific inference.

If they are given by a choice of language or are looked up in a manual of priors, as default Bayesians recommend, then while they are in some sense impersonal, their relevance for predicting and understanding empirical phenomena is unclear. Frequentist methods (significance tests and confidence intervals) are ascertainable at least for a cluster of problems, but how are their long-run error rates relevant in the case of appraising the evidence for a particular scientific inference?

As an outgrowth of number 2, we may identify a third requirement:

3. *The methodology should not be in conflict with intuitions about inductive inference or science or evidence.*

Moreover, it must have a principled, and not an ad hoc, way to avoid any counterintuitive results.

This is a slippery business but it cannot be avoided. Satisfying intuitions about induction and evidence clearly depends on the intended aims of the tools. For example, in appraising Carnapian attempts to arrive at a priori inductive logics there was an appeal to 'inductive intuition'. On this ground, those Carnapian c-functions that result in no learning from positive instances are discredited. Another intuitive principle might be to 'use all relevant evidence', even though notions of relevance differ. But there are far murkier areas where inductive intuitions are unclear, or are intimately tied to background philosophical theories.

Right away we are confronted with issues that depend upon contrasting 'philosophical theories'. Because of this, a philosophical scrutiny may be guilty of imposing its own philosophy on the interpretation of methods. Thus, the threat of circularity looms large in embarking on our mission. Without a separate defense of the philosophical theory that underlies one's foundational scrutiny, there is a danger that the philosophical scrutiny will be question-begging, as often occurs.

Thus, in saying that we recognize the role of intuition in the philosophical scrutiny of methods, we do not mean that there is no justification for them. Quite the opposite: by unearthing these intuitions we can subject them to scrutiny as well.

4.2 How Might Philosophers Construe Ascertainability?

In proposing as a first criterion 'ascertainability', my requirements seem to contradict a still-common manner by which philosophers set out accounts of inductive inference (once called *theories of confirmation*, now known as work in *formal epistemology*). Some view the task in a manner analogous to that of deductive logic. Just as deductive logic tells us that if certain premises are true, then conclusion H follows with certainty, inductive logic would tell us that if certain premises are true, then conclusion H follows with probability. The latter logic is assumed to be well modeled by the probability calculus. As Kyburg puts it (1993, 150), neo-Bayesianism is "yet another effort to convert induction to deduction" in the form of a deductive calculus of probabilities. According to Howson and Urbach (1989, 272),

> "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."

4.2.1 Beyond Validity to Soundness

Howson and Urbach's view of the task of a philosophical account of inductive inference contrasts with what is sought by an account of ampliative inference, or learning from data. To begin with, an adequate account needs to provide guidance for accepting the evidence. In Bayesian philosophy of science especially, the evidence statement is not restricted to a specific statistical model (not that accepting its adequacy is trivial either). Second, since accepting the evidence is not itself a probabilistic inference—it is accepted flat out—at the very least a (non-Bayesian) account of acceptance is needed. Moreover, an ampliative account, at least as I shall view it, requires guidance in detaching claims, whatever form it is to take. Probability theory is deductive all right, but in reducing statistical inference to an application of probability theory, we are missing the inductive component.

Even the strictest deductivist must still wish to apply the valid logical arguments, to obtain ones that are sound or approximately so. This requires affirming premises as at least approximately true, and appealing to methods that, while error-prone, are at least capable of reliably detecting and correcting errors.

4.2.2 Ascertaining Probabilities in the Philosophy and History of Science

I am not claiming that it is the business of the philosopher to tell us how to apply the methods, rather, that it is his or her business to characterize the methods in such a way that scientists could reasonably be supposed to apply them, given limited knowledge in actual contexts. So, for instance, if the method required logical omniscience, it would be a weakness (not necessarily a killing one). Or, if a subjective Bayesian account depended upon elicitations based on betting scenarios, and these were found problematic in science, that would be a weakness in the ascertainability department.

On the other hand, if philosophers of science are proposing the methodology as a way to appraise the rationality of scientific episodes, then *they* should be able to apply it. To illustrate, here is one of the issues that has arisen with respect to the 'problem of old evidence' in determining subjective probabilities: If known evidence is given probability one, then evidence cannot raise the probability to a hypothesis. To avoid this, Bayesian philosophers propose to subtract out the evidence itself from the background, leading Glymour to wonder if this would require philosophers of science to have studied the history of science deeply enough "to make a judgment as to what their degrees of belief would have been in relevant historical periods" (Glymour 1980, 91). To claim merely to be reconstructing the views of scientists, on the other hand, would rob the philosophical account of any normative force.

5. Bayesian Epistemology, Probability 'Logics', and Statistical Science

How shall we understand the meeting ground between statistical science and current work in so-called formal epistemology? Is it inapplicable here?

Analytic epistemology has always limited itself to conceptual analysis of what it means to believe or know various claims, and the formal epistemologist may see him or herself merely as replacing the traditional 'Agent *S* believes that *H*' with a probabilistic rendering, e.g., '*S* assigns a high degree of probability in *H*'. Does it follow then that the formal epistemologist is absolved from taking account of the applicability of the formal methods? If probability is being used to refer to the probability calculus, then I will argue that the answer is no; and I want to devote this section to this question. While it is one that is likely to be of interest mainly to philosophers of science, it is too important to a large segment of current work on probabilistic inference by philosophers to overlook.

5.1 Using Probability and Statistics in Philosophy of Science

Philosophers of science often use probability on the 'meta-level', as it were.

Suppose, for instance, that one starts out with the plausible notion that evidence e confirms a claim H if one believes (or ought to believe?) H more strongly given e than prior to being given evidence e, and assumes that conditional probability is a way to abbreviate this. Then we get a kind of primitive or a priori claim, i.e., e confirms H iff $P(H|e) > P(H)$. The claim is tautologous but also uninformative. But I do not think contemporary Bayesian epistemologists would readily accept that their work is purely a priori.

Bayesian epistemologists seem to wish to claim that there is a place for appealing to probability and statistics in order to get at overarching principles of evidence and logic, and, further, that these principles are informative (and normative) about evidence and inquiry. I am anxious to agree. For this opens the door to at least one of the shared platforms that I would have them step onto.

5.2 A Principle of Good Experimental Design

Colin Howson, who may be credited with the move back to the logics of induction in the late 1990s, makes it clear that he regards Bayesian reconstructions as informative for science.

> "[Bayes's theorem] tells us that the $P(H|e)$ is sensitive to the proportional degree to which e is explained by H as opposed to any other plausible alternative hypotheses. This expresses a basic principle of good experimental design: it should be very unlikely that the sought effect e can be attributed to any cause other than H itself." (Howson 1997)

This principle is to be captured by the fact that $p(H|e)$ is high to the extent that $p(e|\text{not-}H)$ is low, at least in comparison to $p(e|H)$. $P(e|\text{not-}H)$ may be called the *Bayesian-catchall factor*, not-H being all hypotheses in the denial of H. The experimental principle is fine; the problem is supposing that Bayesian machinery supplies it. First, there is the ascertainability problem: arriving at an assignment for the Bayesian catchall would seem to require knowing the future of science, as Wesley Salmon (1966) puts it.

Second, and most important, even in the best cases (i.e., the model is correct, the alternatives are exhaustive), a low value for $p(e|\text{not-}H)$ does not supply the causal or explanatory claim that Howson seeks, to wit, that it is unlikely that the effect e can be attributed to any cause other than H itself. The correct intuition, on the other hand, is easily shown to be captured by the error-probabilistic computation.

5.2.1 A Minimal Principle of Evidence

To explain, let us abbreviate:

(1) It is (very) unlikely that 'the effect of interest' is caused by something other than H.

A good principle of experimental inference is to regard e as evidence of H only when (or only to the extent that) (1) holds. For suppose that (1) is violated, and it is likely that 'the effect of interest' is caused by something other than H.

To claim that e is evidence of H when it is likely that e is attributable to causes other than H (i.e., when (1) is violated) is to follow an inference method with a high probability of being in error. So this would be a very unreliable rule to follow, and H has scarcely passed a stringent or severe test. But this is precisely the '*minimal principle of evidence*' that is at the heart of error-statistical methods.

I propose to allow that both Bayesian and error-statistical philosophies of science would wish to uphold this principle. This provides a shared meeting ground coextensive with current foundational issues of statistical science. A philosophical appraisal of the two statistical philosophies will turn on how well each can capture and further such intuitively plausible principles of scientific learning. If, as I argue, it turns out that error-statistical methods do a better job of supplying methods to satisfy such evidential principles, then this would be a fundamental advantage of the account. From this perspective, oft-repeated criticisms of frequentist methods appear in a different light. Notably, disagreements between posterior probabilities and p-values turn out to correspond with cases where this minimal principle of evidence is violated! (See Kent Staley's contribution.)

5.2.2 Roles of Randomized Trials

Recognizing that an adequate account must be able to satisfy the minimal principle for evidence illuminates corresponding debates about the roles of methodological procedures, such as randomized clinical trials—of increasing interest to philosophers of science (see Senn 2007).

5.3 Getting beyond a Package of Superficiality (Upshot of Section 5)

If much if not most of the work on probability in philosophy of science comes under formal epistemology, and if this enterprise has no need to meet up with statistical methods and their problems, then my meeting ground might seem not to apply to a large segment of this work. I argue that this is a mistake. Especially ironic about this divorce from practice is that it forfeits a central tool for making progress on debates that formal epistemologists care about. Rather than use statistical ideas to answer questions about the methodology of scientific inference, the Bayesian epistemologist starts out by assuming the intuition or principle, the task then being the 'homework problem' of assigning priors and likelihoods that are in sync with the principle. At times this demands beating a Bayesian analysis into line to fit the intuitive principle, while still not getting at its genuine rationale (e.g., with respect to problems of irrelevant conjunctions, and justifying novelty requirements). "The idea of putting probabilities over hypotheses delivered to philosophy a godsend, and an entire package of superficiality." (Glymour 2010, 334)

It follows that formal epistemologists cannot blithely assume to be producing useful rules for science (even at the meta-level) without considering how to cash them out. Plausible principles of evidence might be supposed to be well captured by a given methodology until one asks if the computational components are ascertainable by statistical methods. In so doing, the features of the methods themselves cannot be ignored; nor can any foundational problems surrounding them. Carving out a statistical science-philosophy of science meeting ground is therefore important to all statistical foundations research.

6. Concluding Remarks

If we are to make progress in resolving decades-old controversies which still shake the foundations of statistics, and go on to tackle new ones, I have claimed, we need to dig (or drill?) not shallowly but deeply, a task that requires both statistical and philosophical acumen.

A place to look to meet directly with a host of current foundational problems are the discussions and reactions to 'unifications', or 'reconciliations', of Bayesian and frequentist methods. The purported unifications represent two kinds of 'meetings': between frequentist and Bayesian methods, but also between statistical methodology and epistemology of science. Upon analysis, the unifications are seen to be at odds with both Bayesian and frequentist goals. Even those who pay obeisance to subjective Bayesianism at a 'philosophical' level admit that the statistical methods that actually are used to find things out take a very different form. What then is the statistical philosophy associated with those methods that serve learning? Ironically many seem prepared to allow that Bayesianism still gets it right for epistemology, even as statistical practice calls for methods more closely aligned with frequentist principles. What I would like the reader to consider is that what is right for epistemology is also what is right for statistical learning in practice.

That is, statistical inference in practice deserves its own epistemology. I have suggested one way to characterize and develop this error-statistical epistemology based on the use of error-statistical methods for assessing and controlling the severity of tests. This statistical philosophy is akin to Cox's view that significance tests give a reliable way to use data to indicate how discrepant (and how concordant) a null hypothesis is from the correct understanding of an aspect of a phenomenon of interest, as modeled statistically (Cox and Mayo 2010). The relevant quantities are in terms of degrees of reliability and precision, and degrees of discordance and accordance, rather than degrees of belief or confirmation of hypotheses. The resulting epistemology, I argue, is also more appropriate for the uses to which philosophers of science put statistical methods: to model scientific inference, solve problems about evidence and inference, and critique methodological principles (metamethodology).

But what matters is not whether those engaged in foundational discussions concur with the statistical philosophy I put forward. It suffices that readers recognize that the current situation presents a predicament in need of philosophical illumination, which is the rationale for these forums. In the late 1970s, Lindley said somewhere that the foundations of statistics were so important that everyone should stop what they were doing and sort them out for a couple of years. The same call might be made in 2011 and beyond.

References

Barnard, G. A. (1972), "The Logic of Statistical Inference (review of Ian Hacking, *The Logic of Statistical Inference*)", *British Journal for the Philosophy of Science* 23, 123–132.

Berger, J. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?", *Statistical Science* 18, 1–12.

— (2006), "The Case for Objective Bayesian Analysis", *Bayesian Analysis* 1(3), 385–402.

Bernardo, J. (1997), "Non-informative Priors Do Not Exist: A Discussion", *Journal of Statistical Planning and Inference* 65, 159–189 (with discussion).

— (2010), “Bayesian Objective Hypothesis Testing”, unpublished paper presented at the conference on “Statistical Science and Philosophy of Science: Where Should They Meet?”, June 21 2010 at the London School of Economics. Slides available at URL: http://www.phil.vt.edu/dmayo/conference_2010/Bernardo%20Objective%20Bayesian %20Hypothesis%20testing%206%2021.pdf [10/5/11].

Cox, D. R. (1958), “Some Problems Connected with Statistical Inference”, *Annals of Mathematical Statistics* 29, 357–372.

— (1977), “The Role of Significance Tests”, *Scandinavian Journal of Statistics* 4, 49–70 (with discussion).

— (2006), *Principles of Statistical Inference*, Cambridge: Cambridge University Press.

— and D. V. Hinkley (1974), *Theoretical Statistics*, London: Chapman and Hall.

— and D. Mayo (2010), “Objectivity and Conditionality in Frequentist Inference”, in: Mayo and Spanos 2010, 276–304.

Fisher, R. A. (1955), “Statistical Methods and Scientific Induction”, *Journal of the Royal Statistical Society*, *Series B (Methodological)* 17, 69–78.

— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.

Fraser, D. A. S. (forthcoming), “Is Bayes Posterior Just Quick and Dirty Confidence?”, *Statistical Science* (with discussion).

Gelman A. and C. Shalizi (2011), “Philosophy and the Practise of Bayesian Statistics”, unpublished paper, available at URL: http://www.stat.columbia.edu/~gelman /research/unpublished/philosophy.pdf [10/5/11].

Ghosh, J., M. Delampady and T. Samanta (2006), *An Introduction to Bayesian Analysis, Theory and Methods*, New York: Springer.

Glymour, C. (1980), *Theory and Evidence*, Princeton: Princeton University Press.

— (2010), “Explanation and Truth”, in Mayo and Spanos 2010, 331–350.

Goldstein, M. (2006), “Subjective Bayesian Analysis: Principles and Practice”, *Bayesian Analysis* 1(3), 403–420.

Godambe, V. and D. Sprott (1971) (eds.), *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston of Canada.

Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge: Cambridge Univ. Press.

Harper, W. and C. Hooker (1976) (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. 2, Dordrecht: D. Reidel.

Howson, C. (1997), “A Logic of Induction”, *Philosophy of Science* 64, 268–90.

— and P. Urbach (1993[1989]), *Scientific Reasoning: The Bayesian Approach*, 2nd edn., La Salle: Open Court.

Jeffreys, H. (1961[1939]), *Theory of Probability*, 3rd edn., Oxford: Oxford Univ. Press.

Kadane, J. (2008), “Comment on Article by Gelman”, *Bayesian Analysis* 3(3), 455–458.

Kass, R. E. and L. Wasserman (1996), “The Selection of Prior Distributions by Formal Rules”, *Journal of the American Statistical Association* 91, 1343–1370.

Kyburg, H. E., Jr. (1993), “The Scope of Bayesian Reasoning”, in: Hull, D., M. Forbes and K. Okruhlik (eds.), *PSA 1992*, Proceedings of the 1992 Meeting of the Philosophy of Science Association, Vol. II, East Lansing: Philosophy of Science Association, 139–152.

Lindley, D. V. (1997), "Unified Frequentist and Bayesian Testing of a Precise Hypothesis: Comment", *Statistical Science* 12, 149–152.

Mayo, D. (1985), "Behavioristic, Evidentialist, and Learning Models of Statistical Testing", *Philosophy of Science* 52, 493–516.

— (1992), "Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?", *Synthese* 90, 233–262.

— (1996), *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

— and D. Cox (2010), "Frequentist Statistics as a Theory of Inductive Inference", in: Mayo and Spanos 2011, as reprinted from Mayo and Cox 2006, 247–275.

— and A. Spanos (2006), "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction", *British Journal of Philosophy of Science* 57, 323–357.

— and — (2010) (eds.), *Error and Inference. Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, Chicago: Chicago University Press.

— and — (2011), "Error-Statistics", in: Gabbay, D., P. Thagard and J. Woods (eds.), *Philosophy of Statistics, Handbook of Philosophy of Science*, Elsevier, 152–198.

Neyman, J. (1956), "Note on an Article by Sir Ronald Fisher", *Journal of the Royal Statistical Society*, *Series B (Methodological)* 18, 288–294.

Pearson, E. S. (1955), "Statistical Concepts in Their Relation to Reality", *Journal of the Royal Statistical Society*, *Series B (Methodological)* 17, 204–207.

Peirce, C. S. (1931–35), *The Collected Papers of Charles Sanders Peirce*, vol. 1–6, ed. by C. Hartsthorne and P. Weiss, Cambridge: Harvard University Press.

Popper, K. (1962), *Conjectures and Refutations: The Growth of Scientific Knowledge*, New York: Basic Books.

Salmon, W. (1966), *The Foundations of Scientific Inference*, Pittsburgh: University of Pittsburgh Press.

Senn, S. (2007), *Statistical Issues in Drug Development*, 2nd edn., West Sussex: John Wiley & Sons Inc.

Spanos, A. (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge: Cambridge University Press.

Wasserman, L. (2008), "Comment on Article by Gelman", *Bayesian Analysis* 3(3), 463–466.