RMM Vol. 3, 2012, 71–107
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 Mayo

Statistical Science Meets Philosophy of Science Part 2: Shallow versus Deep Explorations

Abstract:

Inability to clearly defend against the criticisms of frequentist methods has turned many a frequentist away from venturing into foundational battlegrounds. Conceding the distorted perspectives drawn from overly literal and radical expositions of what Fisher, Neyman, and Pearson 'really thought', some deny they matter to current practice. The goal of this paper is not merely to call attention to the howlers that pass as legitimate criticisms of frequentist error statistics, but also to sketch the main lines of an alternative statistical philosophy within which to better articulate the roles and value of frequentist tools.

1. Comedy Hour at the Bayesian Retreat

Overheard at the comedy hour at the Bayesian retreat: Did you hear the one about the frequentist...

> "who defended the reliability of his radiation reading, despite using a broken radiometer, on the grounds that most of the time he uses one that works, so on average he's pretty reliable?"

or

> "who claimed that observing 'heads' on a biased coin that lands heads with probability .05 is evidence of a statistically significant improvement over the standard treatment of diabetes, on the grounds that such an event occurs with low probability (.05)?"

Such jests may work for an after-dinner laugh, but if it turns out that, despite being retreads of 'straw-men' fallacies, they form the basis of why some statisticians and philosophers reject frequentist methods, then they are not such a laughing matter. But surely the drubbing of frequentist methods could not be based on a collection of howlers, could it? I invite the reader to stay and find out.

If we are to take the criticisms seriously, and put to one side the possibility that they are deliberate distortions of frequentist statistical methods, we need to identify their sources. To this end I consider two interrelated areas around which to organize foundational issues in statistics: (1) the roles of probability in induction and inference, and (2) the nature and goals of statistical inference in science or learning. Frequentist sampling statistics, which I prefer to call 'error statistics', continues to be raked over the coals in the foundational literature, but with little scrutiny of the presuppositions about goals and methods, without which the criticisms lose all force.

First, there is the supposition that an adequate account must assign degrees of probability to hypotheses, an assumption often called *probabilism*. Second, there is the assumption that the main, if not the only, goal of error-statistical methods is to evaluate long-run error rates. Given the wide latitude with which some critics define 'controlling long-run error', it is not surprising to find them arguing that (i) error statisticians approve of silly methods, and/or (ii) rival (e.g., Bayesian) accounts also satisfy error statistical demands. Absent this sleight of hand, Bayesian celebrants would have to go straight to the finale of their entertainment hour: a rousing rendition of 'There's No Theorem Like Bayes's Theorem'.

Never mind that frequentists have responded to these criticisms, they keep popping up (verbatim) in every Bayesian and some non-Bayesian textbooks and articles on philosophical foundations. No wonder that statistician Stephen Senn is inclined to "describe a Bayesian as one who has a reverential awe for all opinions except those of a frequentist statistician" (Senn 2011, 59, this special topic of RMM). Never mind that a correct understanding of the error-statistical demands belies the assumption that any method (with good performance properties in the asymptotic long-run) succeeds in satisfying error-statistical demands.

The difficulty of articulating a statistical philosophy that fully explains the basis for both (i) insisting on error-statistical guarantees, while (ii) avoiding pathological examples in practice, has turned many a frequentist away from venturing into foundational battlegrounds. Some even concede the distorted perspectives drawn from overly literal and radical expositions of what Fisher, Neyman, and Pearson 'really thought'. I regard this as a shallow way to do foundations.

Here is where I view my contribution—as a philosopher of science—to the long-standing debate: not merely to call attention to the howlers that pass as legitimate criticisms of frequentist error statistics, but also to sketch the main lines of an alternative statistical philosophy within which to better articulate the roles and value of frequentist tools. Let me be clear that I do not consider this the only philosophical framework for frequentist statistics—different terminology could do as well. I will consider myself successful if I can provide one way of building, or one standpoint from which to build, a frequentist, error-statistical philosophy. Here I mostly sketch key ingredients and report on updates in a larger, ongoing project.

2. Popperians Are to Frequentists as Carnapians Are to Bayesians

Statisticians do, from time to time, allude to better-known philosophers of science (e.g., Popper). The familiar philosophy/statistics analogy—that Popper is to frequentists as Carnap is to Bayesians—is worth exploring more deeply, most notably the contrast between the popular conception of Popperian falsification and inductive probabilism. Popper himself remarked:

> "In opposition to [the] inductivist attitude, I assert that *C(H,***x***)* must not be interpreted as the degree of corroboration of *H* by **x**, unless **x** reports the results of our sincere efforts to overthrow *H*. The requirement of sincerity cannot be formalized—no more than the inductivist requirement that **x** must represent our total observational knowledge." (Popper 1959, 418, I replace 'e' with '**x**')

In contrast with the more familiar reference to Popperian falsification, and its apparent similarity to statistical significance testing, here we see Popper alluding to failing to reject, or what he called the "corroboration" of hypothesis *H*. Popper chides the inductivist for making it too easy for agreements between data **x** and *H* to count as giving *H* a degree of confirmation.

> "Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) *only if these observations or experiments are severe tests of the theory*—or in other words, only if they result from serious attempts to refute the theory." (Popper 1994, 89)

(Note the similarity to Peirce in Mayo 2011, 87, this special topic of RMM.)

2.1 Severe Tests

Popper did not mean to cash out 'sincerity' psychologically of course, but in some objective manner. Further, high corroboration must be *ascertainable:* 'sincerely trying' to find flaws will not suffice. Although Popper never adequately cashed out his intuition, there is clearly something right in this requirement. It is the gist of an experimental principle presumably accepted by Bayesians and frequentists alike, thereby supplying a minimal basis to philosophically scrutinize different methods. (Mayo 2011, section 2.5, this special topic of RMM)

Error-statistical tests lend themselves to the philosophical standpoint reflected in the severity demand. Pretty clearly, evidence is not being taken seriously in appraising hypothesis *H* if it is predetermined that, even if *H* is false, a way would be found to either obtain, or interpret, data as agreeing with (or 'passing') hypothesis *H*. Here is one of many ways to state this:

> *Severity Requirement (weakest):* An agreement between data **x** and *H* fails to count as evidence for a hypothesis or claim *H* if the test

> would yield (with high probability) so good an agreement even if H is false.

Because such a test procedure had little or no ability to find flaws in H, finding none would scarcely count in H's favor.

2.1.1 Example: Negative Pressure Tests on the Deep Horizon Rig

Did the negative pressure readings provide ample evidence that:

> H_0: leaking gases, if any, were within the bounds of safety (e.g., less than θ_0)?

Not if the rig workers kept decreasing the pressure until H passed, rather than performing a more stringent test (e.g., a so-called 'cement bond log' using acoustics). Such a lowering of the hurdle for passing H_0 made it too easy to pass H_0 even if it was false, i.e., even if in fact:

> H_1: the pressure build-up was in excess of θ_0.

That 'the negative pressure readings were misinterpreted', meant that it was incorrect to construe them as indicating H_0. If such negative readings would be expected, say, 80 percent of the time, even if H_1 is true, then H_0 might be said to have passed a test with only .2 severity. Using Popper's nifty abbreviation, it could be said to have low corroboration, .2. So the *error probability* associated with the inference to H_1 would be .8—clearly high. This is not a posterior probability, but it does just what we want it to do.

2.2 Another Egregious Violation of the Severity Requirement

Too readily interpreting data as agreeing with or fitting hypothesis H is not the only way to violate the severity requirement. Using utterly irrelevant evidence, such as the result of a coin flip to appraise a diabetes treatment, would be another way. In order for data $\mathbf{x}$ to succeed in corroborating H with severity, two things are required: (i) $\mathbf{x}$ must fit H, for an adequate notion of fit, and (ii) the test must have a reasonable probability of finding worse agreement with H, were H false. I have been focusing on (ii) but requirement (i) also falls directly out from error statistical demands. In general, for H to fit $\mathbf{x}$, H would have to make $\mathbf{x}$ more probable than its denial. Coin tossing hypotheses say nothing about hypotheses on diabetes and so they fail the fit requirement. Note how this immediately scotches the second howler in the second opening example.

But note that we can appraise the severity credentials of other accounts by using whatever notion of 'fit' they permit. For example, if a Bayesian method assigns high posterior probability to H given data $\mathbf{x}$, we can appraise how often it would do so even if H is false. That is a main reason I do not want to limit what can count as a purported measure of fit: we may wish to entertain different measures for purposes of criticism.

2.3 The Rationale for Severity is to Find Things Out Reliably

Although the severity requirement reflects a central intuition about evidence, I do not regard it as a primitive: it can be substantiated in terms of the goals of learning. To flout it would not merely permit being wrong with high probability—a long-run behavior rationale. In any particular case, little if anything will have been done to rule out the ways in which data and hypothesis can 'agree', even where the hypothesis is false. The burden of proof on anyone claiming to have evidence for H is to show that the claim is not guilty of at least an egregious lack of severity.

Although one can get considerable mileage even with the weak severity requirement, I would also accept the corresponding positive conception of evidence, which will comprise the full severity principle:

> *Severity Principle (full)*: Data $\mathbf{x}$ provide a good indication of or evidence for hypothesis H (only) to the extent that test T severely passes H with $\mathbf{x}$.

Degree of corroboration is a useful shorthand for the degree of severity with which a claim passes, and may be used as long as the meaning remains clear.

2.4 What Can Be Learned from Popper; What Can Popperians Be Taught?

Interestingly, Popper often crops up as a philosopher to emulate—both by Bayesian and frequentist statisticians. As a philosopher, I am glad to have one of our own taken as useful, but feel I should point out that, despite having the right idea, Popperian logical computations never gave him an adequate way to implement his severity requirement, and I think I know why: Popper once wrote to me that he regretted never having learned mathematical statistics. Were he to have made the 'error probability' turn, today's meeting ground between philosophy of science and statistics would likely look very different, at least for followers of Popper, the 'critical rationalists'.

Consider, for example, Alan Musgrave (1999; 2006). Although he declares that "the critical rationalist owes us a theory of criticism" (2006, 323) this has yet to materialize. Instead, it seems that current-day critical rationalists retain the limitations that emasculated Popper. Notably, they deny that the method they recommend—either to accept or to prefer the hypothesis best-tested so far—is reliable. They are right: the best-tested so far may have been poorly probed. But critical rationalists maintain nevertheless that their account is 'rational'. If asked why, their response is the same as Popper's: 'I know of nothing more rational' than to accept the best-tested hypotheses. It sounds rational enough, but only if the best-tested hypothesis so far is itself well tested (see Mayo 2006; 2010b). So here we see one way in which a philosopher, using methods from statistics, could go back to philosophy and implement an incomplete idea.

On the other hand, statisticians who align themselves with Popper need to show that the methods they favor uphold falsificationist demands: that they are capable of finding claims false, to the extent that they are false; and retaining claims, just to the extent that they have passed severe scrutiny (of ways they can be false). Error probabilistic methods can serve these ends; but it is less clear that Bayesian methods are well-suited for such goals (or if they are, it is not clear they are properly 'Bayesian').

3. Frequentist Error-Statistical Tests

Philosophers often overlook lessons from statistical tests because they seek very general accounts of evidence, not limited to formal statistical contexts. I seek a general account as well. However, the elements of statistical tests offer crucial insights for general aspects of inductive inference in science. Most notably, the entire severity assessment fails to be definable without a context in which the error probabilities can be assessed. We will be in a better position to extrapolate to informal settings by recognizing the crucial role of statistical models in providing such a context.

3.1 Probability in Statistical Models of Experiments

Sir David Cox rightly notes that my focus on the use of probability in statistical inference may slight the fundamental role of frequentist probability in modeling phenomena (informal remarks). My excuse is that the main foundational controversy in statistics has revolved around the use of probability in statistical inference. But I agree that the role of frequentist probability in modelling deserves its own focus (see Mayo 1996, chapter 5).

Neyman (1952) emphasizes that the empirical basis for the use of statistical models of experiments is that there are real experiments that "even if carried out repeatedly with the utmost care to keep conditions constant, yield varying results" (25). He gives as examples: an electrically regulated roulette wheel; a coin-tossing machine in which a coin's initial velocity is constant; the number of disintegrations per minute in a quantity of radioactive matter; the tendency for an organism's properties to vary despite homogeneous breeding; measurements of the concentration of an ingredient in a patient's blood. While we cannot predict the outcome of such experiments, a certain pattern of regularity emerges when applied in even a moderately long series of trials. The pattern of regularity is the relative frequency with which specified results occur. Neyman emphasizes that these regularities are just as 'permanent' as any other law-like phenomena.

One can draw out the testable implications of a conjectured model of a phenomenon in science using statistical models that are distinct from substantive scientific ones. We may call the former the experimental, or testable, statistical model, in relation to some substantive model. Often, even without a substantive model or theory—as in the particular case of a so-called exploratory analysis—much can be learned via lower level statistical models of experiment. One strategy is to deliberately introduce probabilistic elements into the data generation so that experimental observations might be framed within statistical models.

For example, if measurements on a patient's blood, when appropriately taken, may be regarded as observing n random variables from a normal distribution with mean equal to μ, $f(\mathbf{x};\mu)$, then we may use experimental results to estimate μ and/or probe various hypotheses about μ's value. Although we construct the model and the experiment, given we have done so, the distribution objectively follows. Capitalizing on knowing how to run real random experiments that correspond appropriately to a mathematically defined probability model, we can deliberately alter the experiment in order to hone our skills at unearthing flaws should we fail to adequately satisfy the statistical model.

Although these models are regarded only as an approximate or idealized representation of the underlying data-generating process, they work because (i) their adequacy for the experiment at hand may be checked by distinct tests, and (ii) they need only capture rather coarse properties of the phenomena being probed (e.g., the relative frequencies of events need to be close to those computed under the statistical models).

3.2 Statistical Test Ingredients

(A) *Hypotheses*. A statistical hypothesis H, generally couched in terms of an unknown parameter θ, is a claim about some aspect of the process that generated the data, $\mathbf{x} = (x_1, \ldots, x_n)$ given in some model of the process. Statistical hypotheses assign probabilities to various outcomes 'computed under the supposition that H_i is correct (about the generating mechanism)'. That is how one should read $f(\mathbf{x}; H_i)$.

Note that this is not a conditional probability, since that would assume that there is a prior probability for H_i. For simplicity I retain this notation where a Bayesian calculation is being considered.

(B) *Distance function*. A function of the data $d(\mathbf{X})$, the *test statistic*, reflects how well or poorly the data $\mathbf{x} = (x_1, \ldots, x_n)$ fit the hypothesis H—the larger the value of $d(\mathbf{x})$ the farther the outcome is from what is expected under H in the direction of alternatives to H, with respect to the particular question being asked. In standard null hypothesis tests, the key is being able to ascertain the probability of different values of $d(\mathbf{X})$ under a test or null hypothesis H_0, and under alternatives. By calculating the probability of outcomes under hypotheses about parameter μ, we can calculate the probabilities of values of statistic d under hypotheses about μ.

(C) *Test rule T*. One type of test procedure might be to infer that $\mathbf{x}$ is evidence of a discrepancy γ from a null hypothesis H_0 just in case $\{d(\mathbf{X}) > c\}$. Thanks to (B), we can calculate the probability of $\{d(\mathbf{X}) > c\}$ under the assumption that H_0 is adequate, as well as under various discrepancies from H_0 contained in the compound alternative H_1. Therefore we can calculate the probability of inferring evidence for discrepancies from H_0 erroneously. Note that such an error probability is given by the probability distribution of $d(\mathbf{X})$—called its *sam-*

pling distribution—computed under one or another hypothesis. I have stated elements (A)–(C) in a generic form, to link with formulations of statistical tests as they typically arise in discussions of foundations. However, to develop an account adequate for solving foundational problems, special stipulations and even reinterpretations of standard notions may be required. The next two elements, (D) and (E), reflect some of these.

(D) The sampling distribution may be used to characterize the capability of the inferential rule to unearth flaws and distinguish hypotheses. At any rate, that is the thesis of the account I aim to develop. What makes an account 'error statistical' is its consideration of these error probabilities. But these computations must be directed by the goal of assessing severity in relation to the particular inference of interest. Not just any use of a sampling distribution makes the account properly 'error statistical'.

(E) *Empirical assumptions*. Quite a lot of empirical background knowledge goes into implementing these computations. We can place them into two groups of questions:

1. How probative would the test be in regard to a particular question if its assumptions were approximately satisfied?
2. Are its assumptions approximately satisfied?

The task of checking assumptions calls for its own discussion (see especially Spanos 2011, this special topic of RMM). To claim that frequentist methods deny the use of background knowledge is absurd. While critics repeatedly warn that this is the consequence of signing up for frequentist statistics, what they mean is that, except for very special cases, we do not use prior probability distributions of unknown parameters (be they degrees of belief or default priors). But Bayesians have not shown that the general kind of background needed is well captured by trying to construct a prior probability distribution of statistical hypothesis H_{i}.

3.3 Hypotheses and Events

In a typical statistical context the hypotheses H_i range over different values of a statistical parameter θ. In the normal distribution example above, θ would be two-dimensional, comprising both the mean and the standard deviation (μ, δ). Since θ 'governs' the distribution, hypothesized values of θ yield probability assignments to the different outcomes $\mathbf{x}$.

A confusion, often lurking in the background of some foundational discussions, stems from mistaking the goal of assigning probabilities to the occurrence of events for that of assigning probabilities to the hypotheses themselves. In inferring a statistical hypothesis H, one is inferring a claim that assigns probabilities to the various experimental outcomes and to the events described in terms of them. This is very different from assigning a probability to H itself (which would speak of the probability of the probabilistic assignment in H).

3.4 Hypotheses Inferred Need Not Be Predesignated

I find it useful to retain the testing language to emphasize the necessary requirement for having evidence, but one need not. Even so, it must not be supposed that we are limited to a rather hackneyed notion of hypotheses tests wherein the hypotheses to be appraised are predesignated as if one were required to know in advance all the possibly interesting inferences that could result from an inquiry. Even where the focus is on statistical tests with prespecified null hypotheses and perhaps directional alternatives, it is a mistake to suppose that this limits the hypotheses whose well-testedness will be of interest. Granted, there is a cluster of canonical null hypotheses, with corresponding methods of analysis. Their chief value, from the current perspective, is to evaluate various discrepancies that are or are not well indicated once the data are in hand. For an excellent taxonomy of types of nulls and corresponding questions see Cox (1977).

This relates to a central contrast between error-statistical and Bayesian methods in the category of 'ascertainability': while the former lets us get started with a battery of simple questions posed by one or more null hypotheses in our repertoire (and corresponding sampling distributions of $d(\boldsymbol{X})$), the latter requires setting out all of the alternative hypotheses that are to be considered. "Full-dress Bayesians", as I. J. Good called them, require, in addition to priors in an exhaustive set of hypotheses, an assignment of utilities or loss functions for decision making. I once invoked a fashion analogy: "Much like ready-to-wear [versus designer] clothes, these 'off the shelf' methods do not require collecting vast resources before you can get going with them." (1996, 100)

Moreover, we wish to distinguish statistical inference from making decisions based on what is warranted to infer. Yet some critics assert, without argument, that frequentist methods and the error-statistical notions based on them are discredited because everyone knows that what we really want are methods for action. "A notion of a severe test without a notion of a loss function is a diversion from the main job of science." (Ziliak and McCloskey 2008, 147) But if one does not first obtain a warranted scientific inference, any subsequent appraisal of expected loss will lack grounding. The politicization of science in the arena of risk assessment is well known, as is the tendency of some policy proponents to regard evidence in support of rival policies 'junk science'. However, if policymaking is inextricably bound up with policy preferences and loss, as Ziliac and McCloskey allege, then appealing to evidence is in danger of becoming just so much window dressing—it is all policy, and evidence-based controversies are merely value-laden disagreements about policy preferences.[1]

Of course, this is a very old view, whether it is called social relativism, postmodernism, or something else (see Mayo 1991; Mayo and Spanos 2006).

[1] Ironically they also fall into misinterpretations of concepts of significance tests that result in supporting the erroneous inferences and fallacies they wish to curtail. See http://www.errorstatistics.com.

4. Neyman's Inferential Side: Neyman on Carnap

Jerzy Neyman, with his penchant for 'inductive behavior' rather than inductive inference, is often seen as a villain in philosophy of statistics disputes. So let me mention a paper of his I came across in a dusty attic not too many years ago with the tantalizing title of "The Problem of Inductive Inference" (Neyman 1955). It is of interest for two reasons: First it reports on the (literal) meeting of a founding frequentist statistician and the philosopher Carnap, in conversation about frequentist inference. In particular, Neyman brings up an erroneous construal of frequentist statistics still common among philosophers. Second, it reveals a use of statistical tests which is strikingly different from the long-run behavior construal most associated with Neyman:

> "When Professor Carnap criticizes some attitudes which he represents as consistent with my ('frequentist') point of view, I readily join him in his criticism without, however, accepting the responsibility for the criticized paragraphs." (13)

4.1 Frequentist Statistics Is Not the Frequentist 'Straight Rule'

Carnap's depiction of 'Neyman's frequentist' is unfortunately still with us. It views frequentists as following a version of the 'inductive straight rule'. Having observed 150 aces out of 1,000 throws with this die, with "no other results of throws with this die [being] known to me" (14), the frequentist infers that "there is a high probability, with respect to the evidence, for the prediction that the relative frequency of aces in a long series of future throws with this die will lie in an interval around 0.15" (ibid.).

Neyman stresses that this overlooks the fact that an "application of any theory of inductive inference can be made only on the ground of a theoretical model of some phenomena, not on the ground of the phenomena themselves" (16). Given the adequacy of a statistical model of experiment—here, the Binomial model—it is possible to use observed relative frequencies to estimate and test claims about the population probability, but it is impossible to do so within Carnap's model-free depiction. What is more, appeals to ignorance, "principles of indifference", are anathema to the solid grounds demanded to vouch for the use of a statistical model. It is still common, however, to hear philosophers depict frequentist statistics as little more than a version of the Carnapian straight rule. No wonder problems of 'the reference class' are pointed to as grounds for criticizing the frequentist approach (e.g., Howson and Urbach 1993; Sober 2008). Within a statistical model, by contrast, the modeler is constrained to an appropriate statistic, here, the sample mean.

4.2 Post-Data Uses of Power

Most interestingly, Neyman continues,

> "I am concerned with the term 'degree of confirmation' introduced by Carnap. [...] We have seen that the application of the locally best one-sided test to the data [...] failed to reject the hypothesis [that the n observations come from a source in which the null hypothesis is true]. The question is: does this result 'confirm' the hypothesis that H_0 is true of the particular data set?" (40–41)

Neyman continues:

> "The answer [...] depends very much on the exact meaning given to the words 'confirmation', 'confidence', etc. If one uses these words to describe one's intuitive feeling of confidence in the hypothesis tested H_0, then [...] the attitude described is dangerous. [...] [T]he chance of detecting the presence [of discrepancy from the null], when only [n] observations are available, is extremely slim, even if [the discrepancy is present]. Therefore, the failure of the test to reject H_0 cannot be reasonably considered as anything like a confirmation of H_0. The situation would have been radically different if the power function [corresponding to a discrepancy of interest] were, for example, greater than 0.95." (42)

The general conclusion is that it is a little rash to base one's intuitive confidence in a given hypothesis on the fact that a test failed to reject this hypothesis. A more cautious attitude would be to form one's intuitive opinion only after studying the power function of the test applied.

4.3 One-sided Test T+

Alluding to our drilling-rig example, the parameter value μ_0 could be the mean pressure beyond which it is considered dangerously high. This is an example of what Cox calls an "embedded null hypothesis" (Cox 1977).

Our measurements $\mathbf{X} = (\mathrm{X}_1,\ldots,\mathrm{X}_\mathrm{n})$ are such that each X_i is Normal, $\mathrm{N}(\mu,\sigma^2)$, (NIID), σ assumed known; and there is a one-sided test T+:

$$H_0\colon \mu \leq \mu_0 \text{ against } H_1\colon \mu > \mu_0.$$

Test statistic $d(\mathbf{X})$ is the sample *standardized* mean, i.e. $d(\mathbf{X}) = (\bar{\mathbf{X}} - \mu)/\sigma_\mathbf{x}$, where $\bar{\mathbf{X}}$ is the sample mean with standard deviation $\sigma_\mathbf{x} = (\sigma/\sqrt{n})$.

The test rule is:

> Infer data $\mathbf{x}$ indicates a (positive) discrepancy from μ_0 iff $\{d(\mathbf{x}) > \mathrm{c}_\alpha\}$.

where c_α is the cutoff corresponding to a difference statistically significant at the α level.

In Neyman's example, the test could not reject the null hypothesis, i.e., $d(\mathbf{x}_0) \leq \mathrm{c}_\alpha$, but (to paraphrase him) the problem is that the chance of detecting the presence of discrepancy γ from the null, with so few observations, is extremely slim, even if γ is present. "One may be confident in the absence of that discrepancy only if the power to detect it were high."

The power of the test T+ to detect γ refers to

(1) $\mathrm{P}(d(\mathbf{X}) > \mathrm{c}_\alpha; \mu = \mu_0 + \gamma)$

It is very interesting to hear Neyman talk this way since it is at odds with the more behavioristic construal he usually championed. Still, power is calculated relative to an outcome just missing the cutoff c_α. This is, in effect, the worst case of a negative (non-significant) result, and if the actual outcome corresponds to an even smaller p-value, that should be taken into account in interpreting the results. It is more informative, therefore, to look at the probability of getting a worse fit (with the null hypothesis) than you did:

(2) $\mathrm{P}(d(\mathbf{X}) > d(\mathbf{x}_0); \mu = \mu_0 + \gamma)$

This gives a measure of the severity (or degree of corroboration) for the inference $\mu < \mu_0 + \gamma$.

Although (1) may be low, (2) may be high.

4.4 Frequentist Principle of Evidence: FEV

The claim in (2) could also be made out viewing the p-value as a random variable, calculating its distribution for various alternatives (Cox 2006, 25). The above reasoning yields a core requirement for frequentist evidence, set out as (FEV) in Mayo and Cox 2010, 256:

> FEV: A moderate p-value is evidence of the absence of a discrepancy γ from H_0 only if there is a high probability the test would have given a worse fit with H_0 (i.e., smaller p-value) were a discrepancy γ to exist.

One must not identify this with what some have called 'post-data power analysis'. (It is beyond the scope of the present discussion.)

It is important to see that it is only in the case of a negative result that severity for various inferences is in the same direction as power. In the case of significant results, where $d(\mathbf{x})$ is in excess of the cutoff, the opposite concern arises—namely, the test is too sensitive. So severity is always relative to the particular inference being entertained: speaking of the 'severity of a test' *simpliciter* is an incomplete statement in this account. These assessments enable sidestepping classic fallacies of tests that are either too sensitive or not sensitive enough. I return to the related 'large n problem' in *section 6.1.3*.

4.5 Pragmatism without Subjectivism

Neyman disparaged "the common element of all writings on the inductive reasoning [that] appears to indicate the conviction of the authors that it is possible to devise a formula of universal validity which can serve as a normative regulator of our beliefs" (Neyman 1957, 15). Instead he offers rules with different performance characteristics and the user is free to choose the one that fits her case best. While this latitude is often the basis of criticisms of error-statistical methods, to Neyman and Pearson this was a central advantage. Still, it must be admitted, that aside from hints and examples, neither he nor Pearson spelled out an overarching logic for using these methods in drawing inferences. That is what my analysis is intended to provide, be it in terms of FEV (for formal statistical contexts) or SEV (for formal and informal assessments).

5. The Error-Statistical Philosophy

I recommend moving away, once and for all, from the idea that frequentists must 'sign up' for either Neyman and Pearson, or Fisherian paradigms. As a philosopher of statistics I am prepared to admit to *supplying* the tools with an interpretation and an associated philosophy of inference. I am not concerned to prove this is what any of the founders 'really meant'.

Fisherian simple-significance tests, with their single null hypothesis and at most an idea of the a directional alternative (and a corresponding notion of the 'sensitivity' of a test), are commonly distinguished from Neyman and Pearson tests, where the null and alternative exhaust the parameter space, and the corresponding notion of power is explicit. On the interpretation of tests that I am proposing, these are just two of the various types of testing contexts appropriate for different questions of interest. My use of a distinct term, 'error statistics', frees us from the bogeymen and bogeywomen often associated with 'classical' statistics, and it is to be hoped that that term is shelved. (Even 'sampling theory', technically correct, does not seem to represent the key point: the sampling distribution matters in order to evaluate error probabilities, and thereby assess corroboration or severity associated with claims of interest.) Nor do I see that my comments turn on whether one replaces frequencies with 'propensities' (whatever they are).

5.1 Error (Probability) Statistics

What is key on the statistics side is that the probabilities refer to the distribution of a statistic $d(\mathbf{X})$—the so-called sampling distribution. This alone is at odds with Bayesian methods where consideration of outcomes other than the one observed is disallowed (likelihood principle [LP]), at least once the data are available.

> "Neyman-Pearson hypothesis testing violates the likelihood principle, because the event either happens or does not; and hence has probability one or zero." (Kadane 2011, 439)

The idea of considering, hypothetically, what other outcomes could have occurred in reasoning from the one that did occur seems so obvious in ordinary reasoning that it will strike many, at least those outside of this specialized debate, as bizarre for an account of statistical inference to banish such considerations. And yet, banish them the Bayesian must—at least if she is being coherent. I come back to the likelihood principle in *section 7*.

What is key on the philosophical side is that error probabilities may be used to quantify the probativeness or severity of tests (in relation to a given inference).

The twin goals of probative tests and informative inferences constrain the selection of tests. But however tests are specified, they are open to an after-data scrutiny based on the severity achieved. Tests do not always or automatically give us relevant severity assessments, and I do not claim one will find just this construal in the literature. Because any such severity assessment is relative to the particular 'mistake' being ruled out, it must be qualified in relation to a given inference, and a given testing context. We may write:

> SEV$(T, \mathbf{x}, H)$ to abbreviate 'the severity with which test T passes hypothesis H with data $\mathbf{x}$'.

When the test and data are clear, I may just write SEV(H).

The standpoint of the severe prober, or the severity principle, directs us to obtain error probabilities that are *relevant* to determining well testedness, and this is the key, I maintain, to avoiding counterintuitive inferences which are at the heart of often-repeated comic criticisms. This makes explicit what Neyman and Pearson implicitly hinted at:

> "If properly interpreted we should not describe one [test] as more accurate than another, but according to the problem in hand should recommend this one or that as providing information which is more relevant to the purpose." (Neyman and Pearson 1967, 56–57)

For the vast majority of cases we deal with, satisfying the N-P long-run desiderata leads to a uniquely appropriate test that simultaneously satisfies Cox's (Fisherian) focus on minimally sufficient statistics, and also the severe testing desiderata (Mayo and Cox 2010).

5.2 Philosophy-Laden Criticisms of Frequentist Statistical Methods

What is rarely noticed in foundational discussions is that appraising statistical accounts at the foundational level is 'theory-laden', and in this case the theory is philosophical. A deep as opposed to a shallow critique of such appraisals must therefore unearth the philosophical presuppositions underlying both the criticisms and the plaudits of methods. To avoid question-begging criticisms, the standpoint from which the appraisal is launched must itself be independently defended.

But for many philosophers, in particular, Bayesians, the presumption that inference demands a posterior probability for hypotheses is thought to be so obvious as not to require support. At any rate, the only way to give a generous interpretation of the critics (rather than assume a deliberate misreading of frequentist goals) is to allow that critics are implicitly making assumptions that are at odds with the frequentist statistical philosophy. In particular, the criticisms of frequentist statistical methods assume a certain philosophy about statistical inference (probabilism), often coupled with the allegation that error-statistical methods can only achieve radical behavioristic goals, wherein long-run error rates alone matter.

Criticisms then follow readily, in the form of one or both:

- Error probabilities do not supply posterior probabilities in hypotheses.
- Methods can satisfy long-run error probability demands while giving rise to counterintuitive inferences in particular cases.

I have proposed an alternative philosophy that replaces these tenets with different ones:

- The role of probability in inference is to quantify how reliably or severely claims have been tested.
- The severity principle directs us to the relevant error probabilities; control of long-run error probabilities, while necessary, is not sufficient for good tests.

The following examples will substantiate and flesh out these claims.

5.3 Severity as a 'Metastatistical' Assessment

In calling severity 'metastatistical', I am deliberately calling attention to the fact that the reasoned deliberation it involves cannot simply be equated to formal-quantitative measures, particularly those that arise in recipe-like uses of methods such as significance tests. In applying it, we consider several possible inferences that might be considered of interest. In the example of test $T+$, the data-specific severity evaluation quantifies the extent of the discrepancy (γ) from the null that is (or is not) indicated by data **x** rather than quantifying a 'degree of confirmation' accorded a given hypothesis. Still, if one wants to emphasize a post-data measure one can write:

> SEV($\mu < \bar{\mathbf{X}}_0 + \gamma\sigma_{\boldsymbol{x}}$) to abbreviate: The severity with which a test $T+$ with a result **x** passes the hypothesis:
>
> ($\mu < \bar{\mathbf{X}}_0 + \gamma\sigma_{\boldsymbol{x}}$) with $\sigma_{\boldsymbol{x}}$ abbreviating $\sigma\sqrt{n}$.

One might consider a series of benchmarks or upper severity bounds:

$$\text{SEV}(\mu < \bar{\mathbf{x}}_0 + 0\sigma_{\boldsymbol{x}}) = .5$$
$$\text{SEV}(\mu < \bar{\mathbf{x}}_0 + .5\sigma_{\boldsymbol{x}}) = .7$$

$\mathrm{SEV}(\mu < \bar{\mathbf{x}}_0 + 1\sigma_{\boldsymbol{x}}) = .84$
$\mathrm{SEV}(\mu < \bar{\mathbf{x}}_0 + 1.5\sigma_{\boldsymbol{x}}) = .93$
$\mathrm{SEV}(\mu < \bar{\mathbf{x}}_0 + 1.98\sigma_{\boldsymbol{x}}) = .975$

More generally, one might interpret nonstatistically significant results (i.e., $d(\mathbf{x}) \leq c_\alpha$) in test T+ above in severity terms:

$\mu \leq \bar{\mathbf{X}}_0 + \gamma_\varepsilon(\sigma/\sqrt{n})$ passes the test T+ with severity $(1 - \varepsilon)$,

for any $\mathrm{P}(d(\mathbf{X}) > \gamma_\varepsilon) = \varepsilon$.

It is true that I am here limiting myself to a case where γ is known and we do not worry about other possible 'nuisance parameters'. Here I am doing philosophy of statistics; only once the logic is grasped will the technical extensions be forthcoming.

5.3.1 Severity and Confidence Bounds in the Case of Test T+

It will be noticed that these bounds are identical to the corresponding upper confidence interval bounds for estimating μ. There is a duality relationship between confidence intervals and tests: the confidence interval contains the parameter values that would not be rejected by the given test at the specified level of significance. It follows that the $(1 - \alpha)$ one-sided confidence interval (CI) that corresponds to test T+ is of form:

$\mu > \bar{\mathbf{X}} - c_\alpha(\sigma/\sqrt{n})$.

The corresponding CI, in other words, would *not* be the assertion of the upper bound, as in our interpretation of statistically insignificant results. In particular, the 97.5 percent CI estimator corresponding to test T+ is:

$\mu > \bar{\mathbf{X}} - 1.96(\sigma/\sqrt{n})$.

We were only led to the upper bounds in the midst of a severity interpretation of negative results (see Mayo and Spanos 2006).

Still, applying the severity construal to the application of confidence interval estimation is in sync with the recommendation to consider a series of lower and upper confidence limits, as in Cox (2006). But are not the degrees of severity just another way to say how probable each claim is? No. This would lead to well-known inconsistencies, and gives the wrong logic for 'how well-tested' (or 'corroborated') a claim is.

A classic misinterpretation of an upper confidence interval estimate is based on the following fallacious instantiation of a random variable by its fixed value:

$\mathrm{P}(\mu < (\bar{\mathbf{X}} + 2(\sigma/\sqrt{n})); \mu) = .975$,

observe mean $\bar{\mathbf{x}}$,

therefore, $\mathrm{P}\ (\mu < (\bar{\mathbf{x}} + 2(\sigma/\sqrt{n})); \mu) = .975$.

While this instantiation is fallacious, critics often argue that we just cannot help it. Hacking (1980) attributes this assumption to our tendency toward 'logicism', wherein we assume a logical relationship exists between any data and hypothesis. More specifically, it grows out of the first tenet of the statistical philosophy that is assumed by critics of error statistics, that of *probabilism*.

5.3.2 Severity versus Rubbing Off

The severity construal is different from what I call the 'rubbing off construal' which says: infer from the fact that the procedure is rarely wrong to the assignment of a low probability to its being wrong in the case at hand. This is still dangerously equivocal, since the probability properly attaches to the method not the inference. Nor will it do to merely replace an error probability associated with an inference to H with the phrase 'degree of severity' with which H has passed. The long-run reliability of the rule is a necessary but not a sufficient condition to infer H (with severity).

The reasoning instead is the counterfactual reasoning behind what we agreed was at the heart of an entirely general principle of evidence. Although I chose to couch it within the severity principle, the general frequentist principle of evidence (FEV) or something else could be chosen.

To emphasize another feature of the severity construal, suppose one wishes to entertain the severity associated with the inference:

$$H\colon \mu < (\bar{\mathbf{x}}_0 + 0\sigma_x)$$

on the basis of mean $\bar{\mathbf{x}}_0$ from test $T+$. H passes with low (.5) severity because it is easy, i.e., probable, to have obtained a result that agrees with H as well as this one, even if this claim is false about the underlying data generation procedure. Equivalently, if one were calculating the confidence level associated with the one-sided upper confidence limit $\mu < \bar{\mathbf{x}}$, it would have level .5. Without setting a fixed level, one may apply the severity assessment at a number of benchmarks, to infer which discrepancies are, and which are not, warranted by the particular data set. Knowing what fails to be warranted with severity becomes at least as important as knowing what is: it points in the direction of what may be tried next and of how to improve inquiries.

5.3.3 What's Belief Got to Do with It?

Some philosophers profess not to understand what I could be saying if I am prepared to allow that a hypothesis H has passed a severe test T with $\mathbf{x}$ without also advocating (strong) belief in H. When SEV*(H)* is high there is no problem in saying that $\mathbf{x}$ warrants H, or if one likes, that $\mathbf{x}$ warrants believing H, even though that would not be the direct outcome of a statistical inference. The reason it is unproblematic in the case where SEV*(H)* is high is:

If SEV*(H)* is high, its denial is low, i.e., SEV*(~H)* is low.

But it does not follow that a severity assessment should obey the probability calculus, or be a posterior probability—it should not, and is not.

After all, a test may poorly warrant both a hypothesis H and its denial, violating the probability calculus. That is, SEV*(H)* may be low because its denial was ruled out with severity, i.e., because SEV*(~H)* is high. But Sev*(H)* may also be low because the test is too imprecise to allow us to take the result as good evidence for H.

Even if one wished to retain the idea that degrees of belief correspond to (or are revealed by?) bets an agent is willing to take, that degrees of belief are comparable across different contexts, and all the rest of the classic subjective Bayesian picture, this would still not have shown the relevance of a measure of belief to the objective appraisal of what has been learned from data. Even if I strongly believe a hypothesis, I will need a concept that allows me to express whether or not the test with outcome $\mathbf{x}$ warrants H. That is what a severity assessment would provide. In this respect, a dyed-in-the wool subjective Bayesian could accept the severity construal for science, and still find a home for his personalistic conception.

Critics should also welcome this move because it underscores the basis for many complaints: the strict frequentist formalism alone does not prevent certain counterintuitive inferences. That is why I allowed that a severity assessment is on the metalevel in scrutinizing an inference. Granting that, the error-statistical account based on the severity principle does prevent the counterintuitive inferences that have earned so much fame not only at Bayesian retreats, but throughout the literature.

5.3.4 Tacking Paradox Scotched

In addition to avoiding fallacies within statistics, the severity logic avoids classic problems facing both Bayesian and hypothetical-deductive accounts in philosophy. For example, tacking on an irrelevant conjunct to a well-confirmed hypothesis H seems magically to allow confirmation for some irrelevant conjuncts. Not so in a severity analysis. Suppose the severity for claim H (given test T and data $\mathbf{x}$) is high: i.e., SEV(T, $\mathbf{x}$, H) is high, whereas a claim J is not probed in the least by test T. Then the severity for the conjunction (H & J) is very low, if not minimal.

> If SEV(Test T, data $\mathbf{x}$, claim H) is high, but J is not probed in the least by the experimental test T, then SEV (T, $\mathbf{x}$, *(H & J)*) = very low or minimal.

For example, consider:

> H: GTR and J: Kuru is transmitted through funerary cannibalism,

and let data $\mathbf{x}_0$ be a value of the observed deflection of light in accordance with the general theory of relativity, GTR. The two hypotheses do not refer to the same data models or experimental outcomes, so it would be odd to conjoin them; but if one did, the conjunction gets minimal severity from this particular data set. Note that we distinguish $\mathbf{x}$ severely passing H, and H being severely passed on all evidence in science at a time.

A severity assessment also allows a clear way to distinguish the well-testedness of a portion or variant of a larger theory, as opposed to the full theory. To apply a severity assessment requires exhausting the space of alternatives to any claim to be inferred (i.e., '*H* is false' is a specific denial of *H*). These must be relevant rivals to *H*—they must be at 'the same level' as *H*. For example, if *H* is asking about whether drug Z causes some effect, then a claim at a different ('higher') level might a theory purporting to explain the causal effect. A test that severely passes the former does not allow us to regard the latter as having passed severely. So severity directs us to identify the portion or aspect of a larger theory that passes. We may often need to refine the hypothesis of stated interest so that it is sufficiently local to enable a severity assessment. Background knowledge will clearly play a key role. Nevertheless we learn a lot from determining that we are *not* allowed to regard given claims or theories as passing with severity. I come back to this in the next section (and much more elsewhere, e.g., Mayo 2010a,b).

6. Some Knock-Down Criticisms of Frequentist Error Statistics

With the error-statistical philosophy of inference under our belts, it is easy to run through the classic and allegedly damning criticisms of frequentist error-statistical methods. Open up Bayesian textbooks and you will find, endlessly reprised, the handful of 'counterexamples' and 'paradoxes' that make up the charges leveled against frequentist statistics, after which the Bayesian account is proferred as coming to the rescue. There is nothing about how frequentists have responded to these charges; nor evidence that frequentist theory endorses the applications or interpretations around which these 'chestnuts' revolve.

If frequentist and Bayesian philosophies are to find common ground, this should stop. The value of a generous interpretation of rival views should cut both ways. A key purpose of the forum out of which this paper arises is to encourage reciprocity.

6.1 Fallacies of Rejection

A frequentist error statistical account, based on the notion of severity, accords well with the idea of scientific inquiry as a series of small-scale inquiries into local experimental questions. Many fallacious uses of statistical methods result from supposing that the statistical inference licenses a jump to a substantive claim that is 'on a different level' from the one well probed. Given the familiar refrain that statistical significance is not substantive significance, it may seem surprising how often criticisms of significance tests depend on running the two together!

6.1.1 Statistical Significance is Not Substantive Significance: Different Levels

Consider one of the strongest types of examples that Bayesians adduce. In a coin-tossing experiment, for example, the result of n trials may occur in testing a null hypothesis that the results are merely due to chance. A statistically significant proportion of heads (greater than .5) may be taken as grounds for inferring a real effect. But could not the same set of outcomes also have resulted from testing a null hypothesis that denies ESP? And so, would not the same data set warrant inferring the existence of ESP? If in both cases the data are statistically significant to the same degree, the criticism goes, the error-statistical tester is forced to infer that there is as good a warrant for inferring the existence of ESP as there is to merely inferring a non-chance effect.[2] But this is false. Any subsequent question about the explanation of a non-chance effect, plausible or not, is at a different level from the space of hypotheses about the probability of heads in Bernouilli trials, and thus would demand a distinct analysis. The nature and threats of error in the hypothesis about Harry's ESP differs from those in merely inferring a real effect. The first significance test did not discriminate between different explanations of the effect, even if the effect is real. The severity analysis makes this explicit.

6.1.2 Error-'fixing' Gambits in Model Validation

That a severity analysis always directs us to the relevant alternative (the denial of whatever is to be inferred) also points up fallacies that may occur in testing statistical assumptions.

In a widely used test for independence in a linear regression model, a statistically significant difference from a null hypothesis that asserts the trials are independent may be taken as warranting one of many alternatives that could explain non-independence. For instance, the alternative H_1 might assert that the errors are correlated with their past, expressed as a lag between trials. H_1 now 'fits' the data all right, but since this is just one of many ways to account for the lack of independence, alternative H_1 passes with low severity. This method has no chance of discerning other hypotheses that could also 'explain' the violation of independence. It is one thing to arrive at such an alternative based on the observed discrepancy with the requirement that it be subjected to further tests; it is another to say that this alternative is itself well tested, merely by dint of 'correcting' the misfit. It is noteworthy that Gelman's Bayesian account advocates model checking. I am not familiar enough with its workings to say if it sufficiently highlights this distinction (Gelman 2011, this special topic of RMM; see also Mayo 2013).

[2] Goldstein (2006) alludes to such an example, but his students, who were supposed to give credence to support his construal, did not. He decided his students were at fault.

6.1.3 Significant Results with Overly Sensitive Tests: Large n Problem

A second familiar fallacy of rejection takes evidence of a statistically significant effect as evidence of a greater effect size than is warranted. It is known that with a large enough sample size any discrepancy from a null hypothesis will probably be detected. Some critics take this to show a rejection is no more informative than information on sample size (e.g., Kadane 2011, 438). Fortunately, it is easy to use the observed difference plus the sample size to distinguish discrepancies that are and are not warranted with severity. It is easy to illustrate by reference to our test T+.

With statistically significant results, we evaluate inferences of the form:

$$\mu > \mu_1 \text{ where } \mu_1 = (\mu_0 + \gamma).$$

Throwing out a few numbers may give sufficient warning to those inclined to misinterpret statistically significant differences. Suppose test T+ has hypotheses

$$H_0\text{: } \mu \leq 0 \text{ vs. } H_1\text{: } \mu > 0.$$

Let $\sigma = 1$, $n = 25$, so $\sigma_{\boldsymbol{x}} = (\sigma/\sqrt{n}) = .2$.

In general:

$$\text{SEV}(\mu > \bar{\mathbf{X}} - \delta_\varepsilon(\sigma\sqrt{n})) = 1 - \varepsilon.$$

Let $\bar{\mathbf{X}} = .4$, so it is statistically significant at the .03 level. But look what happens to severity assessments attached to various hypotheses about discrepancies from 0:

$\text{SEV}(\mu > 0) = .97$
$\text{SEV}(\mu > .2) = .84$
$\text{SEV}(\mu > .3) = .7$
$\underline{\text{SEV}(\mu > .4) = .5}$
$\text{SEV}(\mu > .5) = .3$
$\text{SEV}(\mu > .6) = .16$

I have underlined the inference to $\mu > .4$ since it is an especially useful benchmark.

So, clearly a statistically significant result cannot be taken as evidence for just any discrepancy in the alternative region. The severity becomes as low as .5 for an alternative equal to the observed sample mean, and any greater discrepancies are even more poorly tested! Thus, the severity assessment immediately scotches this well-worn fallacy. Keep in mind that the hypotheses entertained here are in the form, not of point values, but of discrepancies as large or larger than μ (for μ, greater than 0).

Oddly, some Bayesian critics (e.g., Howson and Urbach 1993) declare that significance tests instruct us to regard a statistically significant result at a given level as *more* evidence against the null, the larger the sample size; they then turn around and blame the tests for yielding counterintuitive results! Others have followed suit, without acknowledging this correction from long ago (e.g., Sprenger 2012, this special topic of RMM). In fact, it is indicative of *less* of a discrepancy from the null than if it resulted from a smaller sample size. The same point can equivalently be made for a fixed discrepancy from a null value μ_0, still alluding to our one-sided test *T*+. Suppose $\mu_1 = \mu_0 + \gamma$. An α-significant difference with sample size n_1 passes $\mu > \mu_1$ less severely than with n_2 where $n_2 > n_1$ (see Mayo 1981; 1996).

6.2 P-values Conflict with Posterior Probabilities: The Criticism in Statistics

Now we get to criticisms based on presupposing probabilism (in the form of Bayesian posterior probabilities). Assuming that significance tests really secretly aim to supply posterior probabilities to null hypotheses, the well-known fact that a frequentist p-value can differ from a Bayesian posterior in H_0 is presumed to pose a problem for significance testers, if not prove their out and out "unsoundness" (Howson 1997a,b). This becomes the launching-off point for 'conciliatory' methods that escape the problem while inheriting an improved (Bayesian) foundation. What's not to like?

Plenty, it turns out. Consider Jim Berger's valiant attempt to get Fisher, Jeffreys, and Neyman to all agree on testing (Berger 2003). Taking a conflict between p-values and Bayesian posteriors as demonstrating the flaw with p-values, he offers a revision of tests thought to do a better job from both Bayesian and frequentist perspectives. He has us consider the two-sided version of our Normal distribution test H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$. (The difference between p-values and posteriors is far less marked with one-sided tests.) Referring to our example where the parameter measures mean pressure in the drill rig on that fateful day in April 2010, the alternative hypothesis asserts that there is some genuine discrepancy either positive or negative from some value μ_0.

Berger warns that "at least 22%—and typically over 50%—of the corresponding null hypotheses will be true" if we assume that "half of the null hypotheses are initially true", conditional on a 0.05 statistically significant $d(\mathbf{x})$. Berger takes this to show that it is dangerous to "interpret the p-values as error probabilities", but the meaning of 'error probability' has shifted. The danger follows only by assuming that the correct error probability is given by the proportion of true null hypotheses (in a chosen population of nulls), conditional on reaching an outcome significant at or near 0.05 (e.g., .22%, or over 50%). The discrepancy between p-values and posteriors increases with sample size. If $n = 1000$, a result statistically significant at the .05 level yields a posterior of .82 to the null hypothesis! (A statistically significant result has therefore increased the probability in the null!) But why should a frequentist use such a prior? Why should they prefer to report Berger's 'conditional error probabilities' (of 22%, 50%, or 82%)?

6.2.1 Fallaciously Derived Frequentist Priors

Berger's first reply attempts to give the prior a frequentist flavor: It is assumed that there is random sampling from a population of hypotheses, 50% of which are assumed to be true. This serves as the prior probability for H_0. We are then to imagine repeating the current significance test over all of the hypotheses in the pool we have chosen. Using a computer program, Berger describes simulating a long series of tests and records how often H_0 is true given a small p-value. What can it mean to ask how often H_0 is true? It is generally agreed that it is either true or not true about this one universe. But, to quote C. S. Peirce, we are to imagine that "universes are as plentiful as blackberries", and that we can randomly select one from a bag or urn. Then the posterior probability of H_0 (conditional on the observed result) will tell us whether the original assessment is misleading. But which pool of hypotheses should we use? The 'initially true' percentages will vary considerably. Moreover, it is hard to see that we would ever know the proportion of true nulls rather than merely the proportion that thus far has not been rejected by other statistical tests! But the most serious flaw is this: even if we agreed that there was a 50% chance of randomly selecting a true null hypothesis from a given pool of nulls, .5 would still not give the error statistician a frequentist prior probability of the truth of this hypothesis. It would at most give the probability of the event of selecting a hypothesis with property 'true'. (We are back to Carnap's frequentist.) An event is not a statistical hypothesis that assigns probabilities to outcomes.

Nevertheless, this gambit is ubiquitous across the philosophy of statistics literature. It commits the same fallacious instantiation of probabilities:

> 50% of the null hypotheses in a given pool of nulls are true.
> This particular null hypothesis H_0 was randomly selected from this pool.
> Therefore P(H_0 is true) = .5.

I have called this the *fallacy of probabilistic instantiation*.

6.2.2 The Assumption of 'Objective' Bayesian Priors

When pressed, surprisingly, Berger readily disowns the idea of obtaining frequentist priors by sampling from urns of nulls (though he continues to repeat it). He mounts a second reply: error statisticians should use the 'objective' Bayesian prior of 0.5 to the null, the remaining 0.5 probability being spread out over the alternative parameter space. Many take this to be an 'impartial' or 'uninformative' Bayesian prior probability, as recommended by Jeffreys (1939). Far from impartial, the 'spiked concentration of belief in the null' gives high weight to the null and is starkly at odds with the role of null hypotheses in testing. Some claim that 'all nulls are false', the job being to unearth discrepancies from it.

It also leads to a conflict with Bayesian 'credibility interval' reasoning, since *0* is outside the corresponding interval (I come back to this). Far from considering the Bayesian posterior as satisfying its principles, the error-statistical tester would balk at the fact that use of the recommended priors can result in highly significant results often being construed as no evidence against the null—or even evidence for it!

The reason the Bayesian significance tester wishes to start with a fairly high prior to the null is that otherwise its rejection would be merely to claim that a fairly improbable hypothesis has become more improbable (Berger and Sellke 1987, 115). By contrast, it is informative for an error-statistical tester to reject a null hypothesis, even assuming it is not precisely true, because we can learn how false it is.

Other reference Bayesians seem to reject the 'spiked' prior that is at the heart of Berger's recommended frequentist-Bayesian reconciliation, at least of Berger (2003). This includes Jose Bernardo, who began his contribution to our forum with a disavowal of just those reference priors that his fellow default Bayesians have advanced (2010). I continue to seek a clear epistemic warrant for the priors he does recommend. It will not do to bury the entire issue under a decision-theoretic framework that calls for its own epistemic justification. The default Bayesian position on tests seems to be in flux.

6.3 Severity Values Conflict with Posteriors: The Criticism in Philosophy

Philosophers of science have precisely analogous versions of this criticism: error probabilities (associated with inferences to hypotheses) are not posterior probabilities in hypotheses, so they cannot serve in an adequate account of inference. They are exported to launch the analogous indictment of the severity account (e.g., Howson 1997a,b; Achinstein 2001; 2010; 2011). However severely I might wish to say that a hypothesis H has passed a test, the Bayesian critic assigns a sufficiently low prior probability to H so as to yield a low posterior probability in H. But this is still no argument about why this counts in favor of, rather than against, their Bayesian computation as an appropriate assessment of the warrant to be accorded to hypothesis H. In every example, I argue, the case is rather the reverse. Here I want to identify the general flaw in their gambit.

To begin with, in order to use techniques for assigning frequentist probabilities to events, their examples invariably involve 'hypotheses' that consist of asserting that a sample possesses a characteristic, such as 'having a disease' or 'being college ready' or, for that matter, 'being true'. This would not necessarily be problematic if it were not for the fact that their criticism requires shifting the probability to the particular sample selected—for example, Isaac is ready, or this null hypothesis is true. This was, recall, the fallacious probability assignment that we saw in Berger's attempt in 6.2.1.

6.3.1 Achinstein's Epistemic Probabilist

Achinstein (2010, 187) has most recently granted the fallacy . . . for frequentists:

> "My response to the probabilistic fallacy charge is to say that it would be true if the probabilities in question were construed as rela-

> tive frequencies. However, [. . .] I am concerned with epistemic probability."

He is prepared to grant the following instantiations:

> P% of the hypotheses in a given pool of hypotheses are true (or a character holds for p%).
> The particular hypothesis H_i was randomly selected from this pool.
> Therefore, the objective epistemic probability P(H_i is true) = p.

Of course, epistemic probabilists are free to endorse this road to posteriors—this just being a matter of analytic definition. But the consequences speak loudly against the desirability of doing so.

6.3.2 Isaac and College Readiness

An example Achinstein and I have debated (precisely analogous to several that are advanced by Howson, e.g., Howson 1997a,b) concerns a student, Isaac, who has taken a battery of tests and achieved very high scores, *s*, something given to be highly improbable for those who are not college ready. We can write the hypothesis:

> *H*(I): Isaac is college ready.

And let the denial be *H'*:

> *H'*(I): Isaac is not college ready (i.e., he is deficient).

The probability for such good results, given a student is college ready, is extremely high:

> $P(s|H(I))$ is practically 1,

while very low assuming he is not college ready. In one computation, the probability that Isaac would get such high test results, given that he is not college ready, is .05:

> $P(s|H'(I)) = .05$.

But imagine, continues our critic, that Isaac was randomly selected from the population of students in, let us say, Fewready Town—where college readiness is extremely rare, say one out of one thousand. The critic infers that the prior probability of Isaac's college-readiness is therefore .001:

> (*) $P(H(I)) = .001$.

If so, then the posterior probability that Isaac is college ready, given his high test results, would be very low:

$P(H(I)|s)$ is very low,

even though the posterior probability has increased from the prior in (*).

The fallacy here is that although the probability of a randomly selected student taken from high schoolers in Fewready Town is .001, it does not follow that Isaac, the one we happened to select, has a probability of .001 of being college ready (Mayo 1997; 2005, 117). That Achinstein's epistemic probabilist denies this fallacy scarcely speaks in favor of that account.

The example considers only two outcomes: reaching the high scores *s*, or reaching lower scores, ~*s*. Clearly a lower grade ~*s* gives even less evidence of readiness; that is, $P(H'(I)|{\sim}s) > P(H'(I)|s)$. Therefore, whether Isaac scored as high as *s* or lower, ~*s*, Achinstein's epistemic probabilist is justified in having high belief that Isaac is not ready. Even if he claims he is merely blocking evidence for Isaac's readiness, the analysis is open to problems: the probability of Achinstein finding evidence of Isaac's readiness even if in fact he is ready ($H(I)$ is true) is low if not zero. Other Bayesians might interpret things differently, noting that since the posterior for readiness has increased, the test scores provide at least some evidence for $H(I)$—but then the invocation of the example to demonstrate a conflict between a frequentist and Bayesian assessment would seem to largely evaporate.

To push the problem further, suppose that the epistemic probabilist receives a report that Isaac was in fact selected randomly, not from Fewready Town, but from a population where college readiness is common, Fewdeficient Town. The same scores *s* now warrant the assignment of a strong objective epistemic belief in Isaac's readiness (i.e., $H(I)$). A high-school student from Fewready Town would need to have scored quite a bit higher on these same tests than a student selected from Fewdeficient Town for his scores to be considered evidence of his readiness. When we move from hypotheses like 'Isaac is college ready' to scientific generalizations, the difficulty for obtaining epistemic probabilities via his frequentist rule becomes overwhelming.

We need not preclude that $H(I)$ has a legitimate frequentist prior; the frequentist probability that Isaac is college ready might refer to genetic and environmental factors that determine the chance of his deficiency—although I do not have a clue how one might compute it. The main thing is that this probability is not given by the probabilistic instantiation above.

These examples, repeatedly used in criticisms, invariably shift the meaning from one kind of experimental outcome—a randomly selected student has the property 'college ready'—to another—a genetic and environmental 'experiment' concerning Isaac in which the outcomes are ready or not ready.

This also points out the flaw in trying to glean reasons for epistemic belief with just any conception of 'low frequency of error'. If we declared each student from Fewready to be 'unready', we would rarely be wrong, but in each case the 'test' has failed to discriminate the particular student's readiness from his unreadiness. Moreover, were we really interested in the probability that a student randomly selected from a town is college ready, and had the requisite probability model (e.g., Bernouilli), then there would be nothing to stop the frequentist error statistician from inferring the conditional probability. However, there seems to be nothing 'Bayesian' in this relative frequency calculation. Bayesians scarcely have a monopoly on the use of conditional probability!

6.4 Trivial Intervals and Allegations of Unsoundness

Perhaps the most famous, or infamous, criticism of all—based again on the insistence that frequentist error probabilities be interpreted as degrees of belief—concerns interval estimation methods. The allegation does not merely assert that probability *should* enter to provide posterior probabilities—the assumption I called probabilism. It assumes that the frequentist error statistician also shares this goal. Thus, whenever error probabilities, be they p-values or confidence levels, disagree with a favored Bayesian posterior, this is alleged to show that frequentist methods are unsound!

The 'trivial interval' example is developed by supplementing a special case of confidence interval estimation with additional, generally artificial, constraints so that it can happen that a particular 95% confidence interval is known to be correct—a trivial interval. If we know it is true, or so the criticism goes, then to report a .95 rather than a 100% confidence-level is inconsistent! Non-Bayesians, Bernardo warns, "should be subject to some re-education using well known, standard counter-examples such as the fact that conventional 0.95-confidence regions may actually consist of the whole real line" (2008, 453).

I discussed this years ago, using an example from Teddy Seidenfeld (Mayo 1981); Cox addressed it long before: "Viewed as a single statement [the trivial interval] is trivially true, but, on the other hand, viewed as a statement that all parameter values are consistent with the data at a particular level is a strong statement about the limitations of the data." (Cox and Hinkley 1974, 226) With this reading, the criticism evaporates.

Nevertheless, it is still repeated as a knock-down criticism of frequentist confidence intervals. But the criticism assumes, invalidly, that an error probability is to be assigned as a degree of belief in the particular interval that results. In our construal, the trivial interval amounts to saying that no parameter values are ruled out with severity, scarcely a sign that confidence intervals are inconsistent. Even then, specific hypotheses within the interval would be associated with different severity values. Note: by the hypothesis within the confidence interval, I mean that for any parameter value in the interval μ_1, there is an associated claim of the form $\mu \leq \mu_1$ or $\mu > \mu_1$, and one can entertain the severity for each. Alternatively, in some contexts, it can happen that all parameter values are ruled out at a chosen level of severity.

Even though examples adduced to condemn confidence intervals are artificial, moving outside statistics, the situation in which none of the possible values for a parameter can be discriminated is fairly common in science. Then the 'trivial interval' is precisely what we would want to infer, at least viewing the goal as reporting what has passed at a given severity level. The famous red shift experiments on the General Theory of Relativity (GTR) for instance, were determined to be incapable of discriminating between different relativistic theories of gravity—an exceedingly informative result determined only decades after the 1919 experiments.

6.5 Getting Credit (or Blamed) for Something You Didn't Do

Another famous criticism invariably taken as evidence of the frequentist's need for re-education—and readily pulled from the bag of Bayesian jokes carried to Valencia—accuses the frequentist (error-statistical) account of licensing the following:

> *Oil Exec*: Our inference to H: the pressure is at normal levels is highly reliable!
> *Senator*: But you conceded that whenever you were faced with ambiguous readings, you continually lowered the pressure, and that the stringent 'cement bond log' test was entirely skipped.
> *Oil Exec*: We omitted reliable checks on April 20, 2010, but usually we do a better job—I am giving the average!

He might give further details:

> *Oil Exec*: We use a randomizer that most of the time directs us to run the gold-standard check on pressure. But, April 20 just happened to be one of those times we did the non-stringent test; but on average we do ok.

Overall, this 'test' rarely errs, but that is irrelevant to appraising the inference from the actual data on April 20, 2010. To report the average over tests whose outcomes, had they been performed, are unknown, violates the severity criterion. The data easily could have been generated when the pressure level was unacceptably high, therefore *it misinterprets the actual data*. The question is why anyone would saddle the frequentist with such shenanigans on averages? Lest anyone think I am inventing a criticism, here is the most famous statistical instantiation (Cox 1958).

6.6 Two Measuring Instruments with Different Precisions

A single observation **X** is to be made on a normally distributed random variable with unknown mean μ, but the measurement instrument is chosen by a coin flip: with heads we use instrument E' with a known small variance, say 10^{-4}, while with tails, we use E", with a known large variance, say 10^4. The full data indicate whether E' or E" was performed, and the particular value observed, which we can write as **x**' and **x**", respectively.

In applying our test T+ to a null hypothesis, say, $\mu = 0$, the 'same' value of **X** would correspond to a much smaller p-value were it to have come from E' than if it had come from E". Denote the two p-values as p' and p", respectively.

However, or so the criticism proceeds, the error statistician would report the average p-value: .5(p' + p").

But this would give a misleading assessment of the precision and corresponding severity with either measurement! In fact, any time an experiment E is performed, the critic could insist we consider whether we could have done some other test, perhaps a highly imprecise test or a much more precise test or anything in between, and demand that we report whatever average properties they come up with. The error statistician can only shake her head in wonder that this gambit is at the heart of criticisms of frequentist tests. This makes no sense. Yet it is a staple of Bayesian textbooks, and a main reason given for why we must renounce frequentist methods.

But what could lead the critic to suppose the error statistician must average over experiments not even performed? Here is the most generous construal I can think of. Perhaps the critic supposes what is actually a distortion of even the most radical behavioristic construal:

- If you consider outcomes that could have occurred in hypothetical repetitions of this experiment, you must also consider other experiments that were not (but could have been?) run in reasoning from the data observed, and report some kind of frequentist average.

So if you are not prepared to average over any of the imaginary tests the critic wishes to make up, then you cannot consider *any* data set other than the one observed. This, however, would entail no use of error probabilities. This alone should be a sign to the critic that he has misinterpreted the frequentist, but that is not what has happened.

Instead Bayesians argue that if one tries to block the critics' insistence that I average the properties of imaginary experiments, then "unfortunately there is a catch" (Ghosh, Delampady and Semanta 2006, 38): I am forced to embrace the strong likelihood principle, which entails that frequentist sampling distributions are irrelevant to inference, once the data are obtained. This is a false dilemma: evaluating error probabilities must always be associated with the model of the experiment I have performed. Thus we conclude that "the 'dilemma' argument is therefore an illusion" (Cox and Mayo 2010). Nevertheless, the critics are right about one thing: if we were led to embrace the LP, all error-statistical principles would have to be renounced. If so, the very idea of reconciling Bayesian and error-statistical inference would appear misguided.

7. Can/Should Bayesian and Error Statistical Philosophies Be Reconciled?

Stephen Senn makes a rather startling but doubtlessly true remark:

> "The late and great George Barnard, through his promotion of the likelihood principle, probably did as much as any statistician in the second half of the last century to undermine the foundations of the

> then dominant Neyman-Pearson framework and hence prepare the way for the complete acceptance of Bayesian ideas that has been predicted will be achieved by the De Finetti-Lindley limit of 2020." (Senn 2008, 459)

Many do view Barnard as having that effect, even though he himself rejected the likelihood principle (LP). One can only imagine Savage's shock at hearing that contemporary Bayesians (save true subjectivists) are lukewarm about the LP! The 2020 prediction could come to pass, only to find Bayesians practicing in bad faith. Kadane, one of the last of the true Savage Bayesians, is left to wonder at what can only be seen as a Pyrrhic victory for Bayesians.

7.1 The (Strong) Likelihood Principle (LP)

Savage defines the LP as follows:

> "According to Bayes's theorem, P($x|\ \mu$) [...] constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and more precisely, if y is the datum of some other experiment, and if it happens that P($x|\ \mu$) and P($y|\ \mu$) are proportional functions of μ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of μ." (Savage 1962, 17)

Berger and Wolpert, in their monograph *The Likelihood Principle* (1988), put their finger on the core issue:

> "The philosophical incompatibility of the LP and the frequentist viewpoint is clear, since the LP deals only with the observed x, while frequentist analyses involve averages over possible observations. [...] Enough direct conflicts have been [...] seen to justify viewing the LP as revolutionary from a frequentist perspective." (Berger and Wolpert 1988, 65–66)

The reason I argued in 1996 that "you cannot be a little bit Bayesian", is that if one is Bayesian enough to accept the LP, one is Bayesian enough to renounce error probabilities.

7.2 Optional Stopping Effect

That error statistics violates the LP is often illustrated by means of the *optional stopping effect*. We can allude to our two-sided test from a Normal distribution with mean μ and standard deviation σ, i.e.,

X_i~N(μ,σ) and we wish to test H_0: $\mu = 0$, vs. H_1: $\mu \neq 0$.

Rather than fix the sample size ahead of time, the rule instructs us:

> *Keep sampling until H is rejected at the .05 level* (i.e., keep sampling until $|\bar{\mathbf{X}}| \geq 1.96\ \sigma/\sqrt{n}$).

With n fixed the type 1 error probability is .05, but with this stopping rule the actual significance level differs from, and will be greater than, .05. In the *Likelihood Principle*, Berger and Wolpert claim that "the need here for involvement of the stopping rule clearly calls the basic frequentist premise into question" (74.2–75). But they are arguing from a statistical philosophy incompatible with the error-statistical philosophy which requires taking into account the relevant error probabilities.

Therefore, to ignore aspects of the data generation that alter error probabilities, leads to erroneous assessments of the well testedness, or severity, of the inferences. Ignoring the stopping rule allows a high or maximal probability of error, thereby violating what Cox and Hinkley call "the weak repeated sampling rule". As Birnbaum (1969, 128) puts it, "the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations". From the error statistical standpoint, ignoring the stopping rule allows inferring that there is evidence for a null hypothesis even though it has passed with a low or even 0 severity.

7.3 The Optional Stopping Effect with (Two-sided) Confidence Intervals

The equivalent stopping rule can be framed in terms of the corresponding 95% confidence interval method:

> *Keep sampling until the 95% confidence interval excludes 0.*

Berger and Wolpert concede that using this stopping rule "has thus succeeded in getting the [Bayesian] conditionalist to perceive that $\mu \neq 0$, and has done so honestly" (80–81). This is a striking admission—especially as the Bayesian credibility interval assigns a probability of .95 to the truth of the interval estimate:

$$\mu = \bar{\mathbf{x}} \pm 1.96(\sigma/\sqrt{n}).$$

Does this lead the authors to renounce the LP? It does not. At least not then. To do so would be to renounce Bayesian coherence. From the perspective of the Bayesian (or likelihoodist), to take the stopping rule into account is tantamount to considering the experimenter's intentions (when to stop), which have no place in appraising data. This overlooks the fact that the error statistician has an entirely objective way to pick up on the stopping rule effect, or anything else that influences error probabilities—namely, in the error-statistical report. Although the choice of stopping rule (as with other test specifications) is determined by the intentions of the experimenter, it does not follow that taking account of its influence is to take account of subjective intentions. The allegation is a *non sequitur*.

One need not allude to optional stopping examples to see that error-statistical methods violate the LP. The analogous problem occurs if one has the null hypothesis and is allowed to search for maximally likely hypotheses (Mayo 1996, chap. 9; Mayo and Kruse 2001; Cox and Hinkley 1974)

7.4 Savage's Sleight of Hand in Defense of the LP

While Savage touts the 'simplicity and freedom' enjoyed by the Bayesian, who may ignore the stopping rule, he clearly is bothered by the untoward implications of doing so. (Armitage notes that "thou shalt be misled" if one is unable to take account of the stopping rule.) In dismissing Armitage's result (as no more possible than a perpetual motion machine), however, Savage switches to a very different case—one where the null and the alternative are both (point) hypotheses that have been fixed before the data, and where the test is restricted to these two preselected values. In this case, it is true, the high probability of error is averted, but it is irrelevant to the context in which the optional stopping problem appears—the two-sided test or corresponding confidence interval. Defenders of the LP often make the identical move to the point against point example (Royall 1997). Shouldn't we trust our intuition in the simple case of point against point, some ask, where upholding the LP does not lead to problems (Berger and Wolpert, 83)? No. In fact, as Barnard (1962, 75) explained (to Savage's surprise, at the 'Savage Forum'), the fact that the alternative hypothesis need not be explicit is what led him to deny the LP in general.

7.5 The Counterrevolution?

But all this happened before the sands began to shift some ten years ago. Nowadays leading default Bayesians have conceded that desirable reference priors force them to consider the statistical model, "leading to violations of basic principles, such as the likelihood principle and the stopping rule principle" (Berger 2006, 394). But it is not enough to describe a certain decision context and loss function in which a Bayesian could take account of the stopping rule. Following our requirement for assessing statistical methods philosophically, we require a principled ground (see Mayo 2011). Similarly Bernardo (2005; 2010) leaves us with a concession (to renounce the LP) but without a philosophical foundation. By contrast, a justification that rests on having numbers agree (with those of the error statistician) lacks a philosophical core.

8. Concluding Remarks: Deep versus Shallow Statistical Philosophy

As I argued in part 1 (2011, this special topic of RMM), the Bayesians have ripped open their foundations for approaches that scarcely work from any standpoint. While many Bayesians regard the default Bayesian paradigm as more promising than any of its contenders, we cannot ignore its being at odds with two fundamental goals of the Bayesian philosophical standpoint: incorporating information via priors, and adhering to the likelihood principle. Berger (2003) rightly points out that arriving at subjective priors, especially in complex cases, also produces coherency violations. But the fact that human limitations may prevent attaining a formal ideal is importantly different from requiring its violation in order to obtain the recommended priors (Cox and Mayo 2010). In their attempt to secure default priors, and different schools have their very own favorites, it appears the default Bayesians have made a mess out of their philosophical foundations (Cox 2006; Kadane 2011). The priors they recommend are not even supposed to be interpreted as measuring beliefs, or even probabilities—they are often improper. Were default prior probabilities to represent background information, then, as subjective Bayesians rightly ask, why do they differ according to the experimental model? Default Bayesians do not agree with each other even with respect to standard methods.

For instance, Bernardo, but not Berger, rejects the spiked prior that leads to pronounced conflicts between frequentist p-values and posteriors. While this enables an agreement on numbers (with frequentists) there is no evidence that the result is either an objective or rational degree of belief (as he intends) or an objective assessment of well-testedness (as our error statistician achieves). Embedding the analysis into a decision-theoretic context with certain recommended loss functions can hide all manner of sins, especially once one moves to cases with multiple parameters (where outputs depend on a choice of ordering of importance of nuisance parameters). The additional latitude for discretionary choices in decision-contexts tends to go against the purported goal of maximizing the contribution of the data in order to unearth 'what can be said' about phenomena under investigation. I invite leading reference Bayesians to step up to the plate and give voice to the philosophy behind the program into which they have led a generation of statisticians: *it appears the emperor has no clothes.*

While leading Bayesians embrace default Bayesianism, even they largely seem to do so in bad faith. Consider Jim Berger:

> "Too often I see people pretending to be subjectivists, and then using weakly informative priors that the objective Bayesian community knows are terrible and will give ridiculous answers; subjectivism is then being used as a shield to hide ignorance. In my own more provocative moments, I claim that the only true subjectivists are the objective Bayesians, because they refuse to use subjectivism as a shield against criticism of sloppy pseudo-Bayesian practice." (Berger 2006, 463)

This hardly seems a recommendation for either type of Bayesian, yet this is what the discussion of foundations tends to look like these days. Note too that the ability to use Bayesian methods to obtain 'ridiculous answers' is not taken as grounds to give up on all of it; whereas, the possibility of ridiculous uses of frequentist methods is invariably taken as a final refutation of the account—even though we are given no evidence that anyone actually commits them!

To echo Stephen Senn (2011, this special topic of RMM) perhaps the only thing these Bayesian disputants agree on, without question, is that frequentist error statistical methods are wrong, even as they continue to be used and developed in new arenas. The basis for this dismissal? If you do not already know you will have guessed: the handful of well-worn, and thoroughly refuted, howlers from 50 years ago, delineated in *section 5*.

Still, having found the Bayesian foundations in shambles, even having discarded the Bayesian's favorite whipping boys, scarcely frees frequentist statisticians from getting beyond the classic caricatures of Fisherian and N-P methods. The truth is that even aside from the distortions due to personality frictions, these caricatures differ greatly from the ways these methods were actually used. Moreover, as stands to reason, the focus was nearly always on theoretical principle and application—not providing an overarching statistical philosophy. They simply did not have a clearly framed statistical philosophy. Indeed, one finds both Neyman and Pearson emphasizing repeatedly that these were tools that could be used in a variety of ways, and what really irked Neyman was the tendency toward a dogmatic adherence to a presumed *a priori* rationale standpoint. How at odds with the subjective Bayesians who tend to advance their account as the only rational way to proceed. Now that Bayesians have stepped off their *a priori* pedestal, it may be hoped that a genuinely deep scrutiny of the frequentist and Bayesian accounts will occur. In some corners of practice it appears that frequentist error statistical foundations are being discovered anew. Perhaps frequentist foundations, never made fully explicit, but at most lying deep below the ocean floor, are being disinterred. While some of the issues have trickled down to the philosophers, by and large we see 'formal epistemology' assuming the traditional justifications for probabilism that have long been questioned or thrown overboard by Bayesian statisticians. The obligation is theirs to either restore or give up on their model of 'rationality'.

References

Achinstein, P. (2001), *The Book of Evidence*, Oxford: Oxford University Press.

— (2010), "Mill's Sins or Mayo's Errors?", in: Mayo and Spanos 2010, 170–188.

— (2011), "Achinstein Replies", in: Morgan, G. (ed.), *Philosophy of Science Matters: The Philosophy of Peter Achinstein*, Oxford: Oxford University Press, 258–98.

Armitage, P. (1975), *Sequential Medical Trials*, 2nd ed., New York: Wiley.

Barnard, G. A. (1962), "Prepared Contribution" and "Discussion", in: Barnard and Cox 1962, 39–49; 75–76.

— and D. R. Cox (1962) (eds.), *The Foundations of Statistical Inference: A Discussion*, London: Methuen.

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

— (2006), “The Case for Objective Bayesian Analysis” and “Rejoinder”, *Bayesian Analysis* 1(3), 385–402; 457–464.

— and T. Sellke (1987), “Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence (with discussion)”, *Journal of the American Statistical Association* 82, 112–122.

— and R. L. Wolpert (1988), *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*, 2$^{\text{nd}}$ ed., Lecture Notes–Monograph Series, Vol. 6, Shanti S. Gupta (series ed.), Hayward–California: Institute of Mathematical Statistics.

Bernardo, J. M. (2005), “Reference Analysis”, in: Dey, D. K. and C. R. Rao (eds.), *Handbook of Statistics 25*, Amsterdam: Elsevier, 17–90.

— (2008), “Comment on Article by Gelman”, *Bayesian Analysis* 3(3), 451–454.

— (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 [10/5/2011].

Birnbaum, A. (1969), “Concepts of Statistical Evidence”, in: Morgenbesser, S., P. Suppes and M. White (eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, New York: St. Martin’s Press, 112–43.

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

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

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

— and C. A. Donnelly (2011), *Principles of Applied Statistics*, Cambridge: Cambridge University Press.

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

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

Gelman, A. (2011), “Induction and Deduction in Bayesian Data Analysis”, *Rationality, Markets and Morals (RMM)* 2, 67–78.

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

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

Good, I. J. (1983), *Good Thinking: The Foundations of Probability and Its Applications*, Minneapolis: University of Minnesota Press.

Hacking, I. (1980), “The Theory of Probable Inference: Neyman, Peirce and Braithwaite”, in: Mellor, D. H. (ed.), *Science, Belief and Behavior: Essays in Honour of R. B. Braithwaite*, Cambridge: Cambridge University Press, 141–160.

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

— (1997b), “Error Probabilities in Error”, *Philosophy of Science* 64, 194.

— and P. Urbach (1993[1989]), *Scientific Reasoning: The Bayesian Approach*, 2$^{\text{nd}}$ ed., La Salle: Open Court.

Jeffreys, H. (1939), *Theory of Probability*, Oxford: Oxford University Press.

Kadane J. (2011), *Principles of Uncertainty*, Boca Raton: Chapman & Hall.

Mayo, D. (1981), "In Defense of the Neyman-Pearson Theory of Confidence Intervals", *Philosophy of Science* 48, 269–280.

— (1991), "Sociological vs. Metascientific Views of Risk Assessment", in: Mayo, D. and R. Hollander (eds.), *Acceptable Evidence: Science and Values in Risk Management*, New York: Oxford University Press, 249–279.

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

— (1997), "Error Statistics and Learning from Error: Making a Virtue of Necessity", in: Darden, L. (ed.), *Supplemental Issue PSA 1996: Symposia Papers, Philosophy of Science* 64, S195–S212.

— (2005), "Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved", in: Achinstein, P. (ed.), *Scientific Evidence*, Baltimore: Johns Hopkins University Press, 95–127.

— (2006), "Critical Rationalism and Its Failure to Withstand Critical Scrutiny", in: Cheyne, C. and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer Series Studies in the History and Philosophy of Science, Springer: The Netherlands, 63–99.

— (2010a), "Error, Severe Testing, and the Growth of Theoretical Knowledge", in: Mayo and Spanos 2010, 28–57.

— (2010b), "Towards Progressive Critical Rationalism: Exchanges with Alan Musgrave", in: Mayo and Spanos 2010, 115–124.

— (2011), "Statistical Science and Philosophy of Science: Where Do/Should They Meet in 2011 (and Beyond)?", *Rationality, Markets and Morals (RMM)* 2, Special Topic: Statistical Science and Philosophy of Science, 79–102.

— (2013), "Comments on A. Gelman and C. Shalizi: 'Philosophy and the Practice of Bayesian Statistics'", *British Journal of Mathematical and Statistical Psychology*, forthcoming.

— 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 M. Kruse (2001), "Principles of Inference and Their Consequences", in: Cornfield, D. and J. Williamson (eds.), *Foundations of Bayesianism*, Dordrecht: Kluwer Academic Publishers, 381–403.

— and A. Spanos (2004), "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science* 71, 1007–1025.

— and — (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.

Musgrave, A. (1999), *Essays on Realism and Rationalism*, Amsterdam–Atlanta: Rodopi B.V.

— (2006), "Responses", in: Cheyne, C. and J. Worrall, *Rationality and Reality: Conversations with Alan Musgrave*, The Netherlands: Springer, 293–334.

— (2010), "Critical Rationalism, Explanation and Severe Tests", in: Mayo and Spanos 2011, 88–112.

Neyman, J. (1952), *Lectures and Conferences on Mathematical Statistics and Probability*, 2$^{\text{nd}}$ ed., Washington: U.S. Department of Agriculture.

— (1955), "The Problem of Inductive Inference", *Communications on Pure and Applied Mathematics* 8(1), 13–45.

— (1957), "Inductive Behavior as a Basic Concept of Philosophy of Science", *Revue de l'Institut International de Statistique* 25, 7–22.

— and E. S. Pearson (1967), *Joint Statistical Papers of J. Neyman and E. S. Pearson*, Berkeley: University of California Press.

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. (1959), *The Logic of Scientific Discovery*, New York: Basic Books.

— (1994), *Realism and the Aim of Science: From the Postscript to the Logic of Scientific Discovery*, Oxford–New York: Routledge.

Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, London: Chapman & Hall.

Savage, L. (1962a), "Subjective Probability and Statistical Prnctice", in: Barnard and Cox 1962, 9–35.

— (1962b), "Discussion on Birnbaum", *Journal of the American Statistical Association* 57, 307–8.

Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference*, Dordrecht: Reidel.

Senn, S. (2008), "Comment on an Article by Gelman", *Bayesian Analysis* 3(3), 459-462.

— (2011), "You May Believe You Are a Bayesian But You Are Probably Wrong", *Rationality, Markets and Morals (RMM)* 2, Special Topic: Statistical Science and Philosophy of Science, 48–66.

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

— (2011), "Foundational Issues in Statistical Modeling: Statistical Model Specification and Validation", *Rationality, Markets and Morals (RMM)* 2, 146–178.

Sprenger, J. (2012), "The Renegade Subjectivist: Jose Bernardo's Objective Bayesianism", *Rationality, Markets and Morals (RMM)* 3, Special Topic: Statistical Science and Philosophy of Science, 1–13.

Sober, E. (2008), *Evidence and Evolution: The Logic Behind the Science*, Cambridge: Cambridge University Press.

Ziliak, S. T. and D. N. McCloskey (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Anne Arbor: The University of Michigan Press.