Error Statistics and Learning From Error: Making a Virtue of Necessity

Deborah G. Mayo†‡
Virginia Tech

The error statistical account of testing uses statistical considerations, not to provide a measure of probability of hypotheses, but to model patterns of irregularity that are useful for controlling, distinguishing, and learning from errors. The aim of this paper is (1) to explain the main points of contrast between the error statistical and the subjective Bayesian approach and (2) to elucidate the key errors that underlie the central objection raised by Colin Howson at our PSA 96 Symposium.

1. Introduction.

> The two main attitudes held to-day towards the theory of probability both result from an attempt to define the probability number scale so that it may readily be put in gear with common processes of rational thought. For one school, the degree of confidence in a proposition . . . provides the basic notion to which the numerical scale should be adjusted. The other school notes how in ordinary life a knowledge of the relative frequency of occurrence of a particular class of events in a series of repetitions has again and again an influence on conduct; it therefore suggests that it is through its link with relative frequency that a numerical probability measure has the most direct meaning for the human mind. (Pearson 1950, 228)

The two main attitudes of which Pearson here speaks correspond to two distinct views of the task of a theory of statistics: the first we may call the *evidential-relation* (E-R) view, and the second, the *error statistical* view. This difference corresponds to fundamental differences in the idea of how probabilistic considerations enter in scientific inference

†Department of Philosophy, Virginia Tech, Blacksburg, VA 24061; mayod@vt.edu.

‡I thank E. L. Lehmann for several important error statistical insights.

Philosophy of Science, 64 (Proceedings) pp. S195–S212. 0031-8248/97/64supp-0019$0.00

and thereby in the goal of philosophy of statistics. Evidential-relationship approaches grew quite naturally from what was traditionally thought to be required by a "logic" of confirmation or induction. Most commonly, such approaches seek quantitative measures of the bearing of evidence on hypotheses. What I call error statistical approaches, in contrast, focus their attention on finding general methods or procedures of testing with certain good properties.

In the E-R view, the task of a theory of statistics is to say, for given evidence and hypotheses, how well evidence confirms or supports hypotheses (whether absolutely or comparatively). In this view, the role of statistics is that of furnishing a set of formal rules or a "logic" relating given evidence to hypotheses. The dominant example of such an approach on the contemporary philosophical scene is based on one or another Bayesian measure of support or confirmation. With the Bayesian approach, what we have learned about a hypothesis H from evidence e is measured by the conditional probability of H given e using Bayes's theorem. The cornerstone of the Bayesian approach is the use of prior probability assignments to hypotheses, generally interpreted as an agent's subjective degrees of belief.

In contrast, the methods and models of classical and Neyman-Pearson statistics (e.g., statistical significance tests, confidence interval methods) are primary examples of error probability approaches. These eschew the use of prior probabilities where these are not based on objective frequencies. Probability enters instead as a way of characterizing the experimental or testing process itself; to express how reliably it discriminates between alternative hypotheses and how well it facilitates learning from error. These probabilistic properties of experimental procedures are *error probabilities*.

Several familiar uses of statistics that we read about daily are based on error statistical methods and models: in polls inferring that the proportion likely to vote for a given candidate equals p% plus or minus some percentage points, in reports of statistically significant differences between treated and control groups, in data analyses in physics, astronomy and elsewhere in order to distinguish "signal" from "noise." Despite the prevalence of error statistical methods in scientific practice, the Bayesian Way has been regarded as the model of choice among philosophers looking to statistical methodology. By and large, philosophers of science who consult philosophers of statistics receive the impression that all but Bayesian statistics is discredited.

Jerzy Neyman (co-developer of Neyman and Pearson methods) expressed surprise at the ardor with which subjectivists attacked Neyman-Pearson tests and confidence interval estimation methods back in the 1970s:

> I feel a degree of amusement when reading an exchange between an authority in 'subjectivistic statistics' and a practicing statistician, more or less to this effect:
> *The Authority:* 'You must not use confidence intervals; they are discredited!'
> *Practicing Statistician:* 'I use confidence intervals because they correspond exactly to certain needs of applied work.' (Neyman 1977, 97)

Neyman's remarks hold true today. The subjective Bayesian is still regarded, in many philosophy of science circles, as "the authority" in statistical inference, and scientists from increasingly diverse fields still regard NP methods as corresponding exactly to their needs.

It may seem surprising, given the current climate in philosophy of science, to find philosophers (still) declaring invalid a standard set of experimental methods, rather than trying to understand or explain why scientists evidently (still) find them so useful. I think it is surprising. In any event, I believe it is time to remedy the situation. A genuinely adequate philosophy of experimental inference will only emerge if it is not at odds with statistical practice in science.

My position is that the error statistical approach is at the heart of the widespread applications of statistical ideas in scientific inquiry, and that it offers a fruitful basis for a philosophy of experimental inference. Although my account builds upon several methods and models from classical and Neyman-Pearson (NP) statistics, it does so in ways that depart sufficiently from what is typically associated with these approaches as to warrant some new label. Nevertheless, I retain the chief feature of Neyman-Pearson methods—the centrality of error probabilities—hence the label "error-statistics."

Colin Howson (this issue) argues that error probabilistic methods are in error. After all, scrutinizing an experimental result by considering the error probabilities of the procedures that produced it is anathema to Bayesians, considering as it does outcomes other than the one actually observed. But this just brings out a key point at which the error statistician is at loggerheads with the Bayesian, and is not an argument that the Bayesian Way offers a better account of experimental learning in science. Granting that a very strict (behavioristic) construal of the NP approach—where all that matters is low error rates in the long run—can seem to license counterintuitive inferences, I have sought to erect an error statistical approach that avoids them. In addition to solving a cluster of problems and misinterpretations, this approach attempts to set out a conception of experimental inquiry in which the error statistical tools provide us with just the tools we need for learning in the face of error.

2. A Fundamental Difference in Aims. The error statistical approach seeks tools that can cope with the necessary limitations, and the inevitable slings and arrows, of actual experimental inquiry. The subjective Bayesian upholds a different standard of virtue. Take the very definition of inductive logic, stated by Howson and Urbach:

> Inductive logic—which is how we regard the subjective Bayesian theory—is the theory of inference from some exogenously given data and prior distribution of belief to a posterior distribution. (1989, 290)

Inductive inference from evidence is a matter of updating one's degree of belief to yield a posterior degree of belief (via Bayes's theorem). Where does one get the prior probabilities and the likelihoods required to apply Bayes's theorem? Howson and Urbach (1989, 273) reply that "we are under no obligation to legislate concerning the methods people adopt for assigning prior probabilities. These are supposed merely to characterise their beliefs subject to the sole constraint of consistency with the probability calculus." What about the grounds for accepting the statements of evidence? Just as with arriving at prior probabilities, the evidence is something you need to start out with:

> The Bayesian theory we are proposing is a theory of inference from data; we say nothing about whether it is correct to accept the data. . . . The Bayesian theory of support is a theory of how the *acceptance as true of some evidential statement* affects your belief in some hypothesis. How you came to accept the truth of the evidence, and whether you are correct in accepting it as true, are matters which, from the point of view of the theory, are simply irrelevant. (Howson and Urbach 1989, 272, emphasis added)

This conception of inductive inference no doubt has its virtues. It has the simplicity and cleanness of a deductive logic, virtues that, admittedly, are absent from the error statistician's view of things.

Error statisticians willingly forgo grand and unified schemes for relating their beliefs, preferring a hodgepodge of methods that let them set sail with the kind of information they actually tend to have. Error statisticians appeal to statistical tools as protection from the many ways they know they can be misled by data as well as by their own beliefs and desires. The value of statistical tools for them is to develop strategies that capitalize on their knowledge of mistakes: strategies for collecting and modeling data, for efficiently checking an assortment of errors, and for communicating results in a form that promotes their scrutiny and their extension by others. Once it is recognized that there is a big difference between our aims, we can agree to disagree with

subjective Bayesians as to what virtues our account of scientific inference should possess. Let me highlight some aspects of the error statistical approach that I regard as virtues.

To begin with, rather than starting its work with evidence or data (as Bayesian and other evidential-relationship accounts do), our error statistical approach includes the task of arriving at data—a task that it recognizes as calling for its own inferences. A second point of contrast is that we do not seek to equate the scientific inference with a direct application of some statistical inference scheme.

For example, philosophers often suppose that to apply NP statistics in philosophy of science, scientific inference must be viewed as a matter of accepting or rejecting hypotheses according to whether outcomes fall in acceptance or rejection regions of NP tests. Finding examples where this kind of automatic "accept-reject rule" distorts scientific inference, it is concluded that NP statistics is inappropriate for building an account of inference in science. This conclusion is unwarranted because it overlooks the ways in which these methods are actually used in science. What I am calling the error statistical account, I believe, reflects these actual uses, and shows what is behind the claim of Neyman's scientist, that these methods correspond precisely to certain needs of applied work.

2.1. A Framework of Inquiry. To get at the use of these methods in science, I propose that experimental inference be understood within a framework of inquiry. You cannot just throw some "evidence" at the error statistician and expect an informative answer to the question of what hypothesis it warrants. But neither does the error statistician need to begin with neat and tidy data to get started. A framework of inquiry incorporates methods of experimental design, data generation, modeling, and testing. For each experimental inquiry we can delineate three types of models: *models of primary scientific hypotheses* (or questions), *models of data,* and *models of experiment.*[1]

A substantive scientific inquiry is to be broken down into one or more local or "topical" hypotheses that make up the *primary questions* or *problems* of separate inquiries.[2] Typically, primary problems take the form of estimating quantities of a model or theory, or of testing hypothesized values of these quantities. These local problems often correspond to questions framed in terms of one or more standard or

1. This is akin to the delineation of a hierarchy of models proposed by Patrick Suppes (1969).

2. The term "topical hypotheses" is coined by Hacking (1992). Like topical creams, they are to be contrasted with deeply penetrating theories.

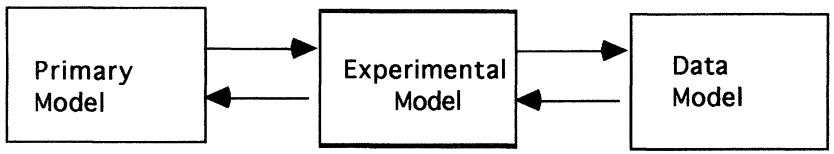

Figure 1: Models of Experimental Inquiry.

canonical errors: about parameter values, about causes, about accidental effects, and about assumptions involved in testing other errors. The experimental models serve as the key linkage models connecting the primary model to the data, links that require, not the raw data itself, but appropriately *modeled data*.

2.2 A Piecemeal Account of Testing. In the error statistical account, formal statistical methods relate to experimental hypotheses, hypotheses framed in the experimental model of a given inquiry. Relating inferences about experimental hypotheses to primary scientific claims is, except in special cases, a distinct step. Yet a third step is called for to link raw data to data models—the real material of experimental inference. So, for example, an inference from data to a primary hypothesis may fail to be warranted either because the experimental inference that is licensed fails to answer the primary question, or it may fail because the assumptions of the experimental data are not met sufficiently by the actual data. In short, there is a sequence of errors that this account directs you to check and utilize along the way, in a kind of piecemeal approach.

Now Howson charges (in unpublished comments for our PSA symposium) that I am "promoting a methodology of piecemeal testing . . . in an attempt to save the game . . . *this is merely making a virtue out of necessity*" (emphasis mine). I accept this charge, for it is a game well worth saving. Where data are inexact, noisy, and incomplete, where extraneous factors are uncontrolled or physically uncontrollable, we simply cannot aspire to the virtues the Bayesian theory demands. We do not have an exhaustive list of hypotheses and probabilities on them, we cannot predict the future course of science, which, to paraphrase Wesley Salmon (1991, 329), would seem to be required to assign the likelihood to the Bayesian catchall factor (i.e., P(e | not-H)). Nor can we, in science, go along with something because individuals strongly believe it, nor can we wait for swamping out of priors to adjudicate disagreements right now about the evidence in front of us.

Nor need we. As a matter of actual fact, we are rather good at

finding things out with a lot less information—*especially* where the threats would be the greatest were we unable to do so. If I cannot actually hold one factor constant to see the effects of others, if I cannot literally manipulate or change, I may still be able to find out *what it would be like* if I could do those things. If I cannot test everything at once, I may be clever enough to test piecemeal. If an event is very rare, I may be able to amplify its effects sufficiently to detect its presence. If I find myself threatened with error, then I need to become a shrewd inquisitor of error. If I cannot face up to these necessary features of experiment by the kind of "white glove" logical analysis of evidence and hypotheses envisioned by E-R approaches, then I am going to have to get down to the nitty-gritty details of the data collection, modeling, and analysis. *Statistics, as I see it, is the conglomeration of systematic tools for carrying out these aims—for making virtues out of necessities.* What is being systematized by these statistical tools is a reflection of familiar, day-to-day learning from errors.

3. Learning From Errors. How do we learn from error? Let me outline in a very sketchy way the kinds of answers that may be found.[3]

1. *After-trial checking (correcting myself).* By "after-trial" I mean after the data or evidence to be used in some inference is at hand. A tentative conclusion may be considered, and we want to check if it is correct. Having made mistakes in reaching a type of inference in the past, we often learn techniques that can be applied the next time to check if we are committing the same error.

In addition to techniques for catching ourselves in error there are techniques for correcting errors. Especially important error-correcting techniques are those designed to go from less accurate to more accurate results, such as taking several measurements and averaging them.

2. *Before-trial planning.* Knowledge of past mistakes gives rise to efforts to avoid the errors ahead of time, before running an experiment or obtaining data. For example, teachers who suspect that knowing the author of a paper may influence their grading may go out of their way to ensure anonymity before starting to grade. This is an informal analogue to techniques of astute experimental design, such as the use of control groups, double-blinding, and large sample size.

3. *An error repertoire.* The history of mistakes made in a type of inquiry gives rise to a list of mistakes to either work to avoid (before-trial planning) or check if committed (after-trial checking), for example, a list of the familiar mistakes when inferring a cause of a correla-

3. A far more complete discussion of these and other aspects of the error statistical approach may be found in Mayo 1996.

tion: Is the correlation spurious? Is it due to an extraneous factor? Am I confusing cause and effect? More homely examples are familiar from past efforts at fixing a car or a computer, or at cooking.

4. *The effects of mistakes.* Through the study of mistakes we learn about the kind and extent of the effect attributable to different errors. This information is then utilized in subsequent inquiries or criticisms. The key is to be able to *discriminate* effects. Perhaps putting in too much water causes the rice to be softer, but not saltier.

Knowledge of the effects of mistakes is often exploited to "subtract out" their influences after the trial. If the effects of different factors can be sufficiently distinguished or subtracted out later, then the inferences are not threatened by a failure to control for them. Thus knowing the effects of mistakes is often the key to justifying inferences.

5. *Simulating errors.* An important way to glean information about the effects of mistakes is by utilizing techniques (real or artificial) to display what it would be like if a given error were committed or a given factor is operative. Such simulations can be used both to rule out and pinpoint factors responsible.

Observing an antibiotic capsule in a glass of water over several days revealed, by the condition of the coating, how an ulceration likely occurred when its coating stuck in my throat. In the same vein, we find scientists appealing to familiar chance mechanisms (e.g., coin tossing) to simulate what would be expected if a result were due to experimental artifacts. Statistical models are valuable because they perform this simulation function formally, by way of (probabilistic) distributions.

6. *Amplifying and listening to error patterns.* One way of learning from error is through techniques for magnifying their effects. I can detect a tiny systematic error in my odometer by driving far enough to a place of known distance. Likewise, a pattern may be gleaned from "noisy" data by introducing a known standard and studying the deviations from that standard. By studying the pattern of discrepancy and by magnifying the effects of distortions, the nature of residuals, and so forth, such deviations can be made to speak volumes.

7. *Robustness.* From all of this information, we also learn when violating certain recommendations or background assumptions does not pose any problem, does not vitiate specific inferences. Such outcomes or inferences are said to be *robust* against such mistakes. An important strategy for checking robustness is to deliberately vary the assumptions and see if the result or argument still holds. This strategy often allows for the argument that the inference is sound, despite violations, that inaccuracies in underlying factors cannot be responsible for a result. For were they responsible, we would not have been able to consistently obtain the same results despite variations.

8. *Severely probing error.* The above seven points form the basis of learning to detect errors. We can put together so potent an arsenal for unearthing a given error that when we fail to find it we have excellent grounds for concluding that the error is absent.

The same kind of reasoning is at the heart of experimental testing. I call it *arguing from error.* After learning enough about certain types of mistakes, we may construct a testing procedure with an overwhelmingly good chance of revealing the presence of a specific error, if it exists—but not otherwise. Such a testing procedure may be called a *severe (or reliable) test,* or a *severe error probe.* If a hypothesized error is not detected by a test that has an overwhelmingly high chance of detecting it, if instead the test yields a result that accords well with no error, then there are grounds for the claim that the error is absent. We can infer something positive, that the particular error is absent (or is no greater than a certain amount). The informal pattern of such an argument from error is guided by the following thesis:

> It is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests taken together) that has a very high probability of detecting an error if (and only if) it exists, nevertheless detects no error.

Its failing to detect the error means it produces a result (or set of results) that accords with the absence of the error. Alternatively, the argument from error can be described in terms of a test of a hypothesis H, that a given error is absent.[4] The evidence indicates the correctness of H when H passes *a severe test*—one with a very high probability of failing H, if H is false. An analogous argument is used to infer the *presence* of an error.

Let me make some remarks on this idea of severity, although I cannot elaborate here in the detail that is merited. First, severity always refers to a particular inference reached or hypothesis passed—a test may be severe for one hypothesis and not another. Second, the statement of "high probability" need not be obtained by reference to a statistical calculation: some of the strongest arguments from error are based on entirely qualitative assessments of severity. This links to the third point, that the formal statement of severity, while a useful summary, is a pale reflection of the actual, real life flesh and blood argument from error. The substantive argument really refers to how ex-

4. In terms of a hypothesis H, the argument from error may be construed as follows: Evidence in accordance with hypothesis H indicates the correctness of H when (and only to the extent that) the evidence results from a procedure that with high probability would have produced a result more discordant from H, were H incorrect.

traordinary the set of circumstances would have to be in order for an error to continually remain hidden from several well-understood detection techniques.

In order to construct severe probes of error, experimental design directs inquiries to be broken down into piecemeal questions. The situation is broken down so that each hypothesis is a local assertion about a particular error in a given experimental framework. It is important to see that the claim that "*H* is false" in assessing severity is not the Bayesian catchall hypothesis. Within an experimental testing model, the falsity of a primary hypothesis *H* takes on a very specific meaning. How to construe it depends on the particular error being ruled out in affirming *H* (e.g., hypothesis *H* asserts a given error is absent, *H* is false asserts that it is present). If *H* states a parameter is greater than some value c, *H* is false states it is less than c; if *H* states that factor x is responsible for at least p% of an effect, its denial states it is responsible for less than p%; if *H* states an effect is caused by factor f—say an artifact of the instrument—*H* is false may say an artifact could not be responsible; if *H* states the effect is systematic, of the sort brought about more often than by chance, then *H* is false states it is due to chance.

This approach lets me test one piece at a time, and there is no subliminal assignment of prior probabilities to the hypotheses that are not being tested by a given test. If I test, seeking to explain your sore throat, whether it is due to strep or not, I am not assigning zero probability to all the other hypotheses that could explain your sore throat. I am just running a test that discriminates strep from no-strep. It is true that to keep alternatives out of a Bayesian appraisal they are effectively assigned a zero probability. That is because the Bayesian appraisal considers a single probability pie, as it were. Appraising any single hypothesis is necessarily a function of all the alternatives in the so-called catchall hypotheses. Lacking this information and desiring to begin to learn something, the scientist, making a virtue of necessity, calls for tools that do not require this information.

4. Do Error Statistical Tests License Unsound Inferences? Howson criticizes a capsulized version of my idea of arguing from error. He refers to it as (*): *e* is a good indication of *H* to the extent that *H* has passed a severe test with e. His argument that (*) is unsound rests on describing a situation in which there is a hypothesis *H* that is indicated according to (*) and yet, intuitively, *e* does not indicate *H*. Both the assumed situation and the intuitions, however, are Bayesian ones. He supposes, in particular, that (i) a certain disease has a very small incidence, say p%, in a given population; and (ii) any randomly chosen test subject

from this population has a (prior) probability of having the disease equal to p. Howson's criticism is that evidence that an error statistician would allegedly take as indicating hypothesis H, the disease is present, yields a very low posterior probability to H—thanks to its low prior probability: "the error probability conditions for a severe test of that particular hypothesis H are clearly satisfied; equally clearly, *passing the test provides no indication of its correctness*" (Howson, this issue, emphasis added).

To begin, it is important to distinguish between questioning the soundness of my rule (*) and questioning the soundness of the formal theory of NP tests, although Howson runs the two together. The error probabilistic calculations of NP tests and confidence intervals are as deductively sound as the Bayesian's calculations. (*), by contrast, is an ampliative and not a deductive rule, and its scrutiny would have to be in contrast to an analogous ampliative Bayesian rule, if only Howson will give us one. Implicitly, he does. Underlying Howson's charge that H's "passing the test provides no indication of its correctness" is something like the following rule:

> *Howson's (implicit) rule:* There is a good indication or strong evidence for the correctness of hypothesis H just to the extent that it has a high posterior probability

which may either be a degree of belief or a relative frequency. We will see that Howson has provided no counterexample to (*)—when it is correctly applied—but rather an illustration of the ways our different rules may conflict.

4.1. The Case With a Frequentist Prior. Before turning to the example, recall that the error statistical account is based upon frequentist methods such as NP tests, and these methods developed precisely for situations in which no frequentist prior is available or even meaningful, as with the majority of scientific hypotheses of interest. Instead, the hypotheses are regarded as unknown constants and only error probabilities *given* one or another hypothesis are considered. The virtue of these methods is their ability to control and learn from these error probabilities without regard to the frequencies with which the hypotheses are true—frequencies that could only make sense where the hypotheses may be regarded as random variables.[5]

But if H is a random variable, and a frequentist prior is available,

5. Although hypotheses regarded as unknown constants are viewed as random variables by Bayesians, Howson is incorrect to maintain that this is also the case for frequentists. It is not (see for example, Neyman 1952, Ch. 1).

the error statistician can use it too.[6] However, Howson erroneously supposes that the probability of randomly selecting a person with the disease from a given population gives a frequentist prior appropriate for the error statistician's question, in the kind of test he describes. It does not.[7] Where the error tester requires positive evidence of having ruled out the presence of disease before issuing a clean bill of health, Howson's analysis *always* concludes that there is no indication that the disease is present!

The disease example. In Howson's example, the *test or null hypothesis* H asserts that a disease is present, and alternative not-H asserts the disease is absent. In this highly artificial example there are only two outputs: an abnormal reading or a normal reading. The test fails to reject H just when an abnormal reading e occurs.[8]

The case of breast cancer screening offers an example where one can find statistics strikingly close to Howson's, and it will help to clarify the intuitions upon which this puzzle rests. The null or test hypothesis is H: breast cancer is present; while "not-H" is that breast cancer is absent. However, "not-H" is a disjunction of hypotheses ranging from the presence of precancerous conditions, to a variety of benign breast diseases, all the way to the absence of breast disease, and in order for us to calculate error probabilities we need to consider specific alternatives under "not-H". We can accommodate Howson's example by focusing on the following null and alternative hypotheses, respectively:

H: breast cancer is present; J: breast disease is absent

In a typical quantitatively modeled test, the null hypothesis H asserts some parameter μ is equal to some value μ_o, and the test rejects H just in case some random variable X is observed to be sufficiently large. Something analogous can be done in modeling the screening for breast cancer. We can imagine that each test involves a set of diagnostic procedures (e.g., mammogram, tumor marker, ultrasound) and X records the number of these that find nothing suspicious. To approximate Howson's example, which supposes that if H is true (breast cancer is present) then an abnormal reading e is assured — i.e., that there is a 0 probability of a Type I error — we can imagine that if even one pro-

6. In such cases, the usual error characteristics of tests can themselves be seen as random variables whose expectations may be assessed. See Neyman 1971, 3.

7. For further discussion of the error involved in this type of example, see Mayo 1997.

8. Howson refers to the two test outputs as "positive" and "negative," where "positive" is the abnormal result, but this may be confusing because a positive test result is commonly equated with rejecting the null. Here, he wants a "positive" (i.e., abnormal) output to *fail* to reject H.

cedure shows suspicion, then the overall result is abnormal, and H is *not* rejected. With today's tests, we could actually get close to Howson's 0 probability of a Type I error, but more realistically, let us suppose:

$$P(e \mid H\text{: breast cancer}) = \text{practically 1.}$$

We are to suppose further that if a person is disease-free, then she very rarely gets an abnormal reading e. Let

$$P(e \mid J) = \text{very low, say .01.}$$

Howson's criticism may be spelled out as follows:

1. An abnormal result e is taken as failing to reject H (i.e., as "accepting H"); while rejecting J, that no breast disease exists.
2. H passes a severe test and thus H is indicated according to (*).
3. But the disease is so rare in the population[9] (from which the patient was randomly sampled) that the posterior probability of H given e is still very low (and that of J given e is still very high).
4. Therefore, "intuitively", H is not indicated but rather J is.
5. Therefore (*) is unsound.

There are two serious problems with this argument: First, premise 2 results from misapplying (*), and second, premise 4 assumes the intuitions of Howson's Bayesian rule. I would deny both premises.

How to Tell the Truth about Failures to Reject H. I grant that there are NP tests that might license the objectionable inference to H, but I deny that they do so severely. As I have already stressed, the error statistician does not use NP tests as automatic accept or reject rules. Rather, one infers those hypotheses that pass severe tests, and to calculate severity correctly requires being clear as to the type and extent of the particular error being probed. Although this is not reducible to a recipe, we can articulate systematic ("metastatistical") rules in order to avoid classic misinterpretations of both rejections of H and failures to reject H in standard testing situations. Howson overlooks these rules and misapplies (*).

Whenever a test result fails to reject null hypothesis H, a standard problem we must be on the lookout for is that the test did not have a good enough chance to reject H even if H is false. We take a lesson

9. Howson's example takes the rate of disease to be .001, but we can consider the population of women in their 30s to arrive at this lower incidence rate of .0005. This allows us to use the more realistic probabiity of P(abnormal result | no disease) = .01 and still get the low posterior probability of H given e (.05) that Howson needs for his argument to go through.

from formal tests of H: Failing to reject the null hypothesis H that $\mu = \mu_0$ does not license the inference that μ is exactly μ_0 because the test may rarely have rejected H even under situations where H is strictly false (i.e., where there are discrepancies from μ_0 in the direction of alternative J). H would only pass with low severity.[10]

The rule for scrutinizing failures to reject H follows the pattern of arguing from error. Let me try to apply it to Howson's example (although without a specific designation of random variable X, this can only be a rough approximation):

> *a.* An abnormal result e is a *poor indication* of the presence of disease more extensive than d if such an abnormal result is probable even with the presence of disease no more extensive than d.
> *b.* An abnormal result e is a *good indication* of the presence of disease as extensive as d if it is very improbable that such an abnormal result would have occurred if a lesser extent of disease were present.

We see that a failure to reject H with e (an abnormal result) does not indicate H so long as there are alternatives to H that would very often produce such a result. For example, the test might often yield a positive result even if only a benign type of breast condition exists. In that case, the severity assessment shows that e is *not* grounds for denying that the condition is of the benign sort: asserting the presence of a disease more serious than a benign one fails to pass a severe test with e. Thus we deny premise 2 of Howson's argument, and notice, we have done so without making use of the prior probability of H. Only discriminations based on error statistical calculations were needed.

Unsoundness? No. A Conflict of Aims? Yes. The second part of the rule, however, allows that the test is warranted in denying hypothesis J: the absence of disease. That is because if J were true, the test would almost surely (99% of the time) *not* have given the abnormal result it did. But cannot Howson's argument be leveled against our allowing "not-J" i.e., *some* disease to severely pass? Let us grant that it can. The argument would go like this: Failing to reject H with e is taken as indicating the denial of J according to (*), but the disease is so rare that the posterior probability of J given e is still very high. Therefore, "intuitively," J is indicated (and denying J is not indicated), and thus (*) gets it wrong. However, this conclusion rests upon Howson's Bayesian intuition, instantiated in premise 4 above.

I readily grant that the Bayesian and the error statistician have very

10. For further discussion of the "rule of acceptance" of H see Mayo 1985, 1989, 1996.

different intuitions here and they stem from the difference in aims sketched earlier. Let us be clear on how an error statistician understands what is being demanded by the test that Howson has specified. (The question of whether it is an appropriate test for some substantive primary question is a distinct problem from the one being confronted just now.) In specifying such a test, with H as the null, and with the 0 or virtually 0 probability of a Type I error, the error tester is saying that we are primarily concerned to avoid giving a woman a false sense of security. We do not want to reject H: the presence of disease, unless we have done an extremely good job *ruling out* the ways in which it may be a mistake to hold that J: a disease-free condition exists. We must do a good job *ruling out* the ways of erroneously inferring J before J is licensed. Finding an abnormal result e clearly does *not* rule out these mistakes, thus e does not warrant J. In other words, in failing to reject H with this test what we mean is: we do not have grounds (of the extent this test is demanding) to assert the absence of breast disease.

But Howson says that we should be primarily concerned to calculate the posterior probability of J given e—which is .95. The appropriate report, he thinks, is that the evidence is a good indication of the absence of the disease.[11] In other words, the Bayesian lab will output a clean bill of health even on the basis of an abnormal reading on the grounds that the incidence of the disease is so low (in the group from which the subject was randomly taken) that the posterior probability is still high that the disease is absent. In the Bayesian screening, no women could ever have breast disease indicated by this test. If the result is normal, the Bayesian infers (with probability approaching 1) there is no breast disease; if the result is abnormal, as we just saw, he also infers no breast disease—although the posterior probability has gone down a little. The Bayesian calculations are correct—the problem is that Howson's lab has not done the job demanded by the error statistical test for breast cancer!

Summary of the Error Statistical Report. Although in practice we are not limited to the artificial "abnormal-normal" dichotomy of Howson's example, even within this limitation we can see how the use of error probabilistic considerations conveys what the results indicate. The abnormal result, one can see, is virtually certain among women

11. *Newsweek* (Feb. 24, 1997, p. 56) recently reported that only 2.5% of women in their 40s who obtain abnormal mammograms are found to have breast cancer. So P(breast cancer | abnormal mammogram) = .025—quite like Howson's made-up example. Using Howson's construal of the evidence, such an abnormal mammogram gives confidence of the *absence* of breast cancer. So the follow-up that discovered these cancers would not have been warranted.

with breast cancer, while it is quite rare, probability .01, among women with no breast disease. Although the abnormal result, we said, did not do a good enough job at discriminating malignant from benign conditions, it clearly did *not* give positive assurance of a disease-free state. In practice, when such a vague index of suspicion of breast cancer results, one of the new highly-sensitive imaging techniques may be indicated to distinguish malignant tumors from various benign breast diseases. But the indication for this further scrutiny hinges upon the soundness of the initial indication—that this result speaks against a disease-free condition.

Of course if there is adequate information on the rates of benign breast disease, the error statistician may report it along with the indication given by the error probabilistic assessments (e.g., "It is very likely that the condition, if any, will prove to be benign"). In practice, however, there is considerable uncertainty as to whether any population from which a given woman is randomly selected provides the appropriate reference class for assessing her particular risk. (Have they considered her age, genetic background, occupation, diet, weight, age of menstruation, etc.?) Of course there are some situations with frequentist priors where the posterior probabilities are the numbers sought, but those cases involve asking about a very different type of error, and (*) gives the right indication for that error (see Note 6).

4.2. The Case With Subjective Priors. Now all this was when the Bayesian uses frequentist prior probabilities and likelihoods. The most troubling problem for the Bayesian account is its use of probabilities construed only as the subjective degrees of belief of a given agent. Although in defending their methods from criticism Bayesians are quick to turn to statistical contexts with frequentist priors (as in the example above), it should not be forgotten that, except for these special cases, the Bayesian obtains the numbers that he says we really want (i.e., posterior probabilities in hypotheses) only by countenancing probabilities understood as subjective degrees of belief. Our analysis of the previous example lets us see how a sufficiently high subjective prior for a hypothesis J countenances high Bayesian confirmation for J, even in the face of evidence that is anomalous for J.

A familiar illustration is found in the subjective Bayesian "solution" to Duhem's problem of where to lay the blame in the face of an anomaly. The situation may parallel the disease example: H entails e whereas e is very improbable given alternative hypothesis J. Prior probabilities again come to J's rescue, now in the form of a high enough prior degree of belief in J. The posterior in J remains high even in the face of anom-

alous result *e*. This "warrants" the scientist in deflecting the anomaly from *J* and instead discrediting rival hypotheses.

To the error statistician, finding an anomaly for *J* hardly counts as having done work to *rule out* the ways in which it can be an error to suppose *J* is correct. The agent's degrees of belief in *J* have nothing to do with it.[12] Indeed, so cavalier a treatment of anomalies can be shown to allow hypothesis *J* to pass with low and even minimal severity. But Bayesians are not required to satisfy these error probabilistic requirements. Howson sees this as a great virtue, heralding the return of common sense.

> Error statisticians . . . have for decades given us something quite different from what we want. . . . Only recently has commonsense returned—commonsense reduced to a calculus . . . now known as the Bayesian theory. (Howson, this issue)

But the Bayesian Way has dominated in philosophy of science for some time. As a result, important aspects of scientific practice are misunderstood or overlooked by philosophers, because these practices reflect error statistical principles that are widespread in science. Howson goes so far as to issue a warning *against* a turn to error statistics: "The message is clear: rely on error probabilities only at your peril" (this issue). But when we really want to learn what our test results are saying, whether about our bodies or about this world, the peril is in relying on the subjective Bayesian screening.

REFERENCES

Hacking, I. (1992), "The Self-Vindication of the Laboratory Sciences", in A. Pickering (ed.), *Science as Practice and Culture*. Chicago: University of Chicago Press, pp. 29–64.

Howson, C. (1997), "Error Probabilities in Error", *Philosophy of Science* 64 (Proceedings): this issue.

Howson, C. and P. Urbach (1989), *Scientific Reasoning: The Bayesian Approach*. La Salle: Open Court.

Mayo, D. (1985), "Increasing Public Participation in Controversies Involving Hazards: The Value of Metastatistical Rules", *Science, Technology, and Human Values* 10: 55–68.

———. (1989), "Toward a More Objective Understanding of the Evidence of Carcinogenic Risk", in A. Fine and J. Leplin (eds.), *PSA 1988*, Vol. 2, East Lansing, MI: Philosophy of Science Association, pp. 489–503.

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

———. (1997), "Duhem's Problem, The Bayesian Way, and Error Statistics, or 'What's Belief Got To Do With It'?" and "Response to Howson and Laudan", *Philosophy of Science* (June 1997), in press.

Neyman, J. (1952), *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed. Washington, DC: U.S. Department of Agriculture.

12. See Mayo 1997.

———. (1971), "Foundations of Behavioristic Statistics", in V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston of Canada, pp.1–13 (comments and reply, pp. 14–19).

———. (1977), "Frequentist Probability and Frequentist Statistics", *Synthese* 36: 97–131.

Newsweek, (February 24, 1997), "The Mammogram War", pp. 54–58.

Pearson, E. S. (1950), "On Questions Raised by the Combination of Tests Based on Discontinuous Distributions", *Biometrika* 37: 383–398, as reprinted in E. S. Pearson (1966), *The Selected Papers of E.S. Pearson*. Berkeley: University of California Press, pp. 217–232.

Salmon, W. (1991), "The Appraisal of Theories: Kuhn Meets Bayes", in A. Fine, M. Forbes, and L. Wessels (eds.), *PSA 1990*, Vol. 2. East Lansing, MI: Philosophy of Science Association, pp. 325–332.

Suppes, P. (1969), "Models of Data", in his *Studies in the Methodology and Foundations of Science*. Dordrecht: D. Reidel, pp. 24–35.