SCIENTIFIC EVIDENCE

Philosophical Theories & Applications

Edited by

PETER ACHINSTEIN

The Johns Hopkins University Press
Baltimore & London

© 2005

6

Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses

Deborah G. Mayo

According to Karl Popper, "Mere supporting instances are as a rule too cheap to be worth having; . . . any support capable of carrying weight can only rest upon ingenious tests, undertaken with the aim of refuting our hypothesis, if it can be refuted" (Popper, 1983: 30). As intuitively plausible as this remark is, philosophical accounts of evidence, including Popper's, have been stymied by the problem of identifying those cases where accordance between data x and a hypothesis H should really count as evidence for H. As increasingly available formal techniques for data mining and model selection bring the price of obtaining "good fitting models" ever lower, the problem is further exacerbated. The latest high-powered computer packages offer a welter of algorithms for "automatically" selecting among this explosion of models, but as each boasts different, and incompatible, selection criteria, we are thrown back to our basic question: What is required to severely discriminate among well-fitting models such that, *when a claim (or hypotheses or model) survives this test the resulting data count as good evidence for the claim's correctness, or dependability, or adequacy*. The question, as I see it, is fundamentally one for philosophy of science, although its adequate answer demands a combination of methodological, empirical, and statistical means.

In this chapter I sketch a conception of evidence as the result of surviving a severe or risky test:

> *Data x in test T provide good evidence for inferring H (just) to the extent that hypothesis H has passed a severe test T with x.*

(I use the word *hypothesis* to cover claims about models as well as the more usual statistical hypotheses.) This discussion is intended as a continuation and extension of the existing program that I call "the error statistical account," being built on ideas from the school of statistical testing based on frequentist error probabilities (e.g., significance levels, confidence levels). On this conception of evidence, even claims that are highly beliefworthy, or highly probable (however probability is construed), would not get credit from tests that poorly probed them; evidential credit goes only to those specific hypotheses or models that have survived probative testing.

I begin by explicating the concepts involved more clearly than before to address several misunderstandings, both of the statistical tools and of the broader account of evidence and inference that I wish to base upon those tools. I then confront a general type of criticism that is repeatedly raised (usually, but not exclusively from Bayesian quarters) which would, if correct, vitiate the account of evidence I am advancing, namely, that *the severity account of evidence is inadequate because hypotheses that pass severe tests on my account may be accorded low posterior degrees of support, probability, or belief.* I consider some of the variants of the criticism raised by philosophers, such as Achinstein and Howson, as well as by statistical practitioners such as Berger, Cohen, and Meehl. The problem revolves around what might be dubbed the "*highly probed versus highly probable debate,*" and although it has been around for some time, progress has been greatly hampered due to the lack of an adequate account of "highly probed" (i.e., of passing a severe test). After clarifying the notion of severity I have in mind, I argue, with respect to each variant of the criticism, that (1) the probabilistic assignment commits a fallacy, which may be called the *fallacy of instantiating probabilities*; and (2) the error statistical assessment of the data, but not the assessment advocated by the critic, is in sync with the goals of severity, as well as our intuitions about when data should count as supporting evidence in science.

The Severity Requirement: Some Examples

College Readiness

If we are testing how well a student, Isaac, has mastered high school material so as to be considered sufficiently ready for work in a four-year college, a test that covers work from 11th and 12th grade science, history, and mathematical problems (in geometry, algebra, trigonometry, and pre-calculus), requires writing a critical essay, and so on, is obviously *more difficult to pass* than one which only requires showing minimal proficiency in these subjects at a 6th or 7th grade level. It would be regarded as more searching, more probing, and more severe. The understanding behind this commonplace judgment is roughly this: Achieving a passing or high score is easier and more likely to have come about with the less severe test than the more severe one, *even among students who have not* *mastered the bulk of high school material, and hence are not "college-ready."* We deny that Isaac's passing score provides good evidence that he fully masters high school material if we learn that even students who mastered little of the material would very probably have scored as high as, or even higher than, he did. In so doing, we are demanding that *before regarding a passing result as genuine evidence for the correctness of a given claim or hypothesis H, it does not suffice to merely survive a test; such survival must be something that is very difficult to achieve if in fact H deviates from what is truly the case.*

By the same token, if the test *is* sufficiently stringent, such that it is practically impossible for students who have not mastered at least p% of high school material to achieve a score as high as Isaac's, then we regard his passing grade as evidence that he has mastered at least this much. The same reasoning abounds in science, as some recent examples illustrate.

Hormone Replacement Therapy

Until very recently, millions of women in the United States were routinely prescribed hormone replacement therapy (HRT) on the grounds that there was excellent evidence that:

> *H*: HRT is advantageous to menopausal and post-menopausal women.

Data, collected over many years, were taken to show HRT's benefits, not only for the effects of menopause, but also for reducing the risk of heart attack and stroke, various cancers, memory loss, and a host of age-related deficiencies—thereby allowing women to remain "Feminine Forever," to cite the title of one famous book.[1] The results of recent, large, randomized treatment-control studies, however, are widely taken to show that such data actually failed to provide evidence for the alleged benefits. Moreover, the new data are regarded as strong evidence that HRT, especially if taken for more than five years, increases the chance of breast cancer and fails to provide the supposed protection against heart disease and memory loss.

"We painted too rosy a picture," the acting director of the Women's Health Initiative conceded.[2] Due to the nature of the retrospective studies on which such sanguine recommendations were based, the previous studies are now known to have had little capacity to distinguish the benefits of HRT from confounding factors. In particular, the women using HRT have been found to have characteristics that are separately correlated with the beneficial outcomes (e.g., they are healthier, have better access to medical care, and are better educated than women not taking HRT).

The underlying reasoning is this: *data x do not supply evidence for a (causal) hypothesis if it would have been just as easy for the observed association between x and the* *hypothesized factor in H to have come about due to factors other than the one claimed in H*. The agreement (between x and H) from the retrospective studies failed to provide evidence in support of H because H did not thereby pass a test that we would consider severe.

The Columbia Space Shuttle

Within hours of the Columbia space shuttle disaster, it was hypothesized that the crash was due to overheating from foam tiles striking the left wing, and yet the available data behind this early conjecture (temperature readings, photographs of displaced tiles during takeoff) was scarcely solid evidence for this hypothesis. Although the "damaged tile on ascent" hypothesis seemed to fit, and could "account for" the data, the early data did not constitute the results of probing numerous other possible fault lines.

Contrast this to the recently completed report by the NAS Board. First, the NAS report tightens up the fit by providing a very detailed version of the "damaged tile on ascent" hypothesis:

> H: The physical cause of the loss of *Columbia* and its crew was a breach in the Thermal Protection System on the leading edge of the left wing, caused by a piece of insulating foam which separated from the left bipod ramp section of the External Tank at 81.7 seconds after launch, and struck the wing in the vicinity of the lower half of Reinforced Carbon-Carbon panel number 8.

Second, the report presents "aerodynamic, thermodynamic, sensor data timeline, debris reconstruction, and imaging evidence—to show that all five independently arrive at the same conclusion," H, which has thereby passed a severe test. *When we evaluate evidence in this way, we are scrutinizing inferences according to the severity of the tests they have survived.* Note that the severity assessment, as is typical, is based not on a single data set but on numerous separate probes taken together.

Severe Tests

From these examples, we can identify what is required for a hypothesis (or model or prediction) H to have passed a good or severe test with data x. Although at minimum a passing result requires a "good agreement" between data x and hypothesis H, more is required. In addition to finding a good fit between x and H, we need to be able to say that the test was really probative—that so good a fit between data x and H is practically impossible or extremely improbable (or an extraordinary coincidence, or the like) if in fact it is a mistake to regard x as evidence for H.

Severity Requirement

So one way we can encapsulate the severity requirement is this:

> *Hypothesis H passes a severe test T with x if (and only if):*
>
> (i) x agrees with or "fits" H (for a suitable notion of fit[3]), and
>
> (ii) test T would (with very high probability) have produced a result that fits H less well than x does, if H were false or incorrect.

Data x provide good evidence for inferring H only if x results from a procedure which, *taken as a whole*, constitutes H having passed a severe test—that is, a procedure which would have (at least with very high probability) unearthed any error or flaw in the inference to H. Far from wishing to justify the familiar inductive rule from observing that n% of A's have been B's in a sample, to an inference that n% of A's are B's in a given population, we can see that such a rule would license inferences that had not passed severe tests, and would be highly unreliable. Likewise we eschew going from correlational data to causal claims, without demonstrating that errors have been well ruled out. Correlational data x provide good evidence for H only when x is generated in a procedure that did a good job of ruling out the ways it can be an error to go from x to H.

Qualitative Assessments of Severity

Although the previous section encapsulates severity in terms of probability, it is not required that a formal probability model be definable to sustain a severe test; in fact, the strongest severity arguments seem to arise from wholly informal appraisals. For instance, were Isaac to continually score high marks, regardless of how difficult or advanced we make the exam (SAT, advanced placement tests, etc.), we are clearly on firm ground in regarding the passing results as excellent evidence that Isaac knows the given material without any recourse to formal models. The term "probability" here may serve merely to pay obeisance to the fact that all empirical claims are strictly fallible, even if a counterexample is never actually instantiated. In engineering and other contexts it is also common to work without a well-specified probability model for catastrophic events, and yet the same requirement about evidence holds. Modifying my definition for such contexts, the engineer Yakov Ben-Haim suggests, "We are subjecting a proposition to a severe test if an erroneous inference concerning the truth of the proposition can result only under extraordinary circumstances" (Ben-Haim, 2001: 214).[4] If one thinks of the circumstances that would have to obtain for the NAS board to be substantially wrong in their interpretation of the data, one gets the idea of what is meant by "extraordinary" here.

Severity Principle

Whether severity is understood quantitatively or qualitatively, in terms of probability or in terms of non-probabilistic, but still formal, notions, the overarching principle of evidence remains. We may refer to it as *the severity principle*:

> *Severity Principle*: Data ***x*** (produced by process G) provide a good indication or evidence for hypothesis *H* (just) to the extent that test *T* severely passes *H* with ***x***.

Within this analysis, hypothesis *H* is regarded (or modeled) as a claim about some aspect of the process that generated the data, G. According to the severity principle, when hypothesis *H* has passed a highly severe test (something that may require several individual tests taken together), we can regard data ***x*** as supplying good grounds that we have ruled out the ways it can be a mistake to regard ***x*** as having been generated by the procedure described by *H*.

Dangerous Misunderstandings

Although a full understanding of the severity principle, and of how to calculate severity, demands careful discussion beyond this paper, the central points I need to make require avoiding some common misunderstandings, especially in regard to the context where the severity requirement refers to a probabilistic model.

A Test (in the current account) Does Not Require Starting With a Hypothesis H

I find it useful to adopt the language of testing because it seems the best way to highlight the challenge (or *agon*) that an inference is required to survive before allowing that there is evidence for it. However, there are common conceptions philosophers typically hold about tests that I wish to deny (C. S. Peirce may be the sole exception). In particular, there is no presumption that the hypothesis is arrived at first (somehow), that is, it is not assumed that the data ***x*** must be "novel" in some sense. Beginning with data ***x*** and appropriately arriving at a hypothesis *H* (which might be a model, or other claim), as in cases of estimation or model searching, may still permit ***x*** to be a severe test for *H*. In fact, I developed the severity notion precisely to distinguish between legitimate and illegitimate cases where data ***x*** are used both to arrive at and test a hypothesis, that is, where a hypothesis *H* is "use-constructed" (see, for example, Mayo, 1991; 1996).

What our analysis demands is recognizing *how* various data-dependent procedures *may* create obstacles for the severity with which given claims may be said to pass (analogous to the way significance levels and other error probabilities are altered by certain selection procedures and stopping rules in statistics).[5] However, so long as the overall analysis satisfies severity requirements, x is good evidence for H (or, equivalently, H has passed a severe test with x).

A Severity Assessment Is Always Relative to the Hypothesis That "Passes"

It is common to talk as if a severity assessment attaches to the test itself—but doing so leads to untoward results. One cannot answer the question: How severe is test *T*? without including the particular inference that is claimed to have passed the test, if any. In other words, whatever claim or hypothesis one is contemplating inferring, on the basis of x, is the claim to regard as *passing* for purposes of appraising severity. The great advantage of relativizing the assessment to the particular inference (and the particular data set) is that high severity is always what is wanted for evidence.[6] No problem occurs unless one forgets that a given test may severely pass one hypothesis and not another, even among the hypotheses under consideration. This confusion most readily takes the form of what might be called *the criticism from overly sensitive tests.*

The Criticism from Overly Sensitive Tests

Severity cannot be a sensible desiderata, so the criticism goes, because a test may be made so severe that even a trivially small departure from a hypothesis H will result in inferring H'—where H' is a rival to H, or an assertion about some anomaly or error in H. What this criticism overlooks is that the inference whose severity we would need to consider in that case is H′; but having put H to a stringent test is not to have stringently probed H'! The misunderstanding behind the criticism boils down to thinking that H′ has passed a severe test, as I am defining it, but in fact it is quite the opposite.

Consider our test for deficiencies in college readiness and the hypothesis:

H: Isaac is college-ready (i.e., not deficient) ,

as against

H': Isaac is not college-ready.

We can make the tests so hard, and the hurdle for regarding grades as evidence for H so high, that his scores are practically always going to lead to denying H and inferring H' (he is deficient). However, H' has passed a test with *very low* severity because it would very often lead to inferring H', even if H' is false and actually H is true.

The Criticism from Overly Sensitive Tests is often given with regard to testing models: Look, if we make the test so severe, the critic says, we are always going to find some flaw in the model—models are always approximate. This not only fails to employ "severity" as I have defined it, but it also overlooks the central job it is designed to perform: (a) Severity lets us show that if we make the test so sensitive, then the assertion "this model is flawed" is *not* going to pass a severe test; and (b) a severity assessment issues in a report of what (if anything) *does* pass severely (e.g., in statistical tests, it reveals just how small the discrepancy indicated is).[7] (In a severity interpretation of statistical tests, (b) is formalized in a "rule for rejection." See Mayo, 1996.)

Severity Condition (ii) Differs from Saying That x Is Very Improbable Given Not-H

That is, condition (ii) is not merely to assert that $P(x;H \text{ is false})$ is low, where "$P(x;H \text{ is false})$" is to be read: "the probability of x under the assumption that H is false."[8] This is called the *likelihood* of H given x. For a familiar example, H_1 might be that a coin is fair, and x the result of n flips. For any x one can construct a hypothesis H_2 that makes the data maximally likely; for example, H_2 can assert that the probability of heads is 1 just on those tosses that yield heads, and 0 otherwise. $P(x;H_1)$ is very low and $P(x;H_2)$ is high, but H_2 has not passed a severe test because one can always construct *some such maximally likely hypothesis or other* to perfectly fit the data on coin tosses, even though it is false and the coin is perfectly fair (i.e., H_1 is true). The test that H_2 passes has minimal severity. (This is a case of what I call "gellerization.")

A remark on notation: I am using ";" in writing $P(x;H)$—in contrast to the notation typically used for a conditional probability, $P(x/H)$—in order to emphasize that severity does *not* use a conditional probability which, strictly speaking, requires that the prior probabilities $P(H_i)$ be well-defined, for an exhaustive set of hypotheses. As we will see, such priors are not well defined in the frequentist severity account within which I am working.

The Degree of Severity with Which a Test Passes H Is Not the Degree of Probability of H

Finding that a hypothesis H severely passes test T with data x does not license a posterior probability assignment to H, a notion which depends on having prior probability assignments to the hypotheses under consideration. Such Bayesian calculations (from one of a number of schools of Bayesianism) are at odds with the severity principle in general. This, of course, is behind the "highly probed versus highly probable" subtitle of this chapter, and it is an issue to be unpacked in detail in later sections.

It should be emphasized that "H is false" is not the so-called *catchall factor*, that is, the disjunction of hypotheses other than H, including those not yet even thought of. Instead, it refers to a specific error that hypothesis H may be seen to be denying. The particular experimental context serves to ensure that H is sufficiently local so that H and its complement exhaust the space of hypotheses for the experiment at hand.

Logics of Evidential Relationship: Subjective Bayesian Philosophy

The severity account of evidence is at odds with familiar philosophical accounts of evidence or confirmation that seek to provide one or more measures of the logical relationship between given evidence (or evidence statements) and hypotheses. A leading example of such a *logic of evidential relationship* (E-R logic) is the subjective Bayesian account, which I understand here to refer to the kind of simple model found in Howson and Urbach (1989; 1993) and in many other philosophical works. "Inductive logic—which is how we regard the subjective Bayesian theory—is a theory of (degree of belief) consistency and thereby also a theory of inference from some exogenously given data and prior distribution of belief to a posterior distribution" (1993: 419). In the Bayesian model, data x would usually be regarded as evidence for H to the extent that the posterior probability in H given x exceeds the prior in H, although various measures of extent of evidence are possible.

The reliance on subjective probabilities in an exhaustive set of hypotheses; the fact that the import of the data comes in only by way of the likelihoods; and that "how you came to accept the truth of the evidence, and whether you are correct in accepting it as true" are regarded as "simply irrelevant" to the account are all reasons that the Bayesian approach is at serious odds with the conception of induction as severe testing (ibid., 1993: 407). Strong degree of belief in a hypothesis H, coupled with believing that the data x are very improbable under alternatives to H, suffices for high Bayesian support for H. It does not suffice for regarding H as having passed a severe test.

Existing quantitative E-R measures of support, confirmation, or the like may be evaluated by means of the severity requirement by regarding each as supplying one or another measures of "fit" (as in condition (i) of severity). However, for any such measure of fit between x and H, the severe tester wants to ask: *How often would so good a fit result (from a series of applications of the test) under the assumption that H is false?* The probability that a hypothesis H would pass a test, under the assumption that H is false, is an *error probability* or error frequency. The severity with which H passes is high just in case (or just to the extent that) this error probability is very low.[9] It is this crucial role of error probabilities that makes it appropriate to call the severity account an error probability account, even where the assessment remains qualitative. Because the error probability assessment demands taking this additional step beyond the Bayesian or other evidential-relation accounts, it is not surprising that the latter fail to control error probabilities, as we have defined them.

Popperian Severity

It was Popper who is best known for insisting on severe tests, and while the severity requirement is clearly Popperian *in spirit,* Popper never adequately captured the severity notion. Although Popper offered various formal definitions that would *potentially* measure the degree to which x corroborates H, $C(H,x)$, he claimed that in order for it to actually measure corroboration, x would have to be the result of a severe test: "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 substitute his *h, e* with x, H for consistency with my notation).

Under "inductivist" Popper includes Bayesians both of Carnapian and subjective varieties, as well as those holding variants of induction by enumeration. The important kernel of rightness here is that these inductive logics of evidential relationship made it too easy to find evidence in support of hypotheses: *Their measures of evidential relationship may be satisfied without satisfying the requirement of severity.* The formal counterpart to this claim is that all such algorithms lack formal niches through which to pick up on aspects of the evidence that are relevant for assessing if the test was really good at probing H's errors, that is, niches for assessing severity in my sense. For example, there are aspects of the generation of data and the specification of hypotheses to test that alter a test's error probabilities and yet do not change the ratio of likelihoods of the hypotheses. Hence, if "same likelihood" means "same evidence" there will be no way to formally pick up on a difference that makes a difference, at least to one who requires severity. Unfortunately, Popper's computations suffered from just this weakness.[10]

Popper's Comparativist Account

According to Popper, data $\boldsymbol{x}$ pass H severely if: H fits or entails data $\boldsymbol{x}$ and $\boldsymbol{x}$ is improbable "without" H or under the assumption that H is false. But his formal measures never got beyond requirements about comparative likelihoods, such as

(a) $P(\boldsymbol{x};H)$ = high (or maximal)

(b) $P(\boldsymbol{x};H$ is false) = low

Moreover, Popper had no machinery to compute requirement (b), but instead supposed (b) is satisfied when

(b)′ $P(\boldsymbol{x};H')$ = low

where H' is the currently best-tested alternative. However, since H and H' need not be, and generally would not be, exhaustive of the space of possible hypotheses, (b)′ certainly does not warrant (b). Data may satisfy Popper's requirement without satisfying severity as I am using that term; the mere fact that $\boldsymbol{x}$ is counterpredicted by the currently best-tested alternative to H does not suffice for severely probing H. Moreover, even (b), as Popper states it, seems to be merely a likelihood, and not an error probability.

A Simple Comparative Likelihood Account

Popper's definition boils down to what may be called *a simple comparative likelihood account*: $\boldsymbol{x}$ is evidence for H if the likelihood of H exceeds the likelihood of alternative H'. (Hacking 1965 had at one time embraced such an account.) Of course this clearly does not do justice to what Popper sought to require, which seems to be the reason he despaired of formally characterizing the conception that he held to be so vital. Popper never saw how error probabilistic notions permit capturing the intended severity requirement rigorously, rather than leaving it at the vague level of the psychological intentions of the tester (to sincerely try to find flaws in H).[11]

An equally, if not more crucial difference between the approach I am putting forward and Popper's is that even where a hypothesis has passed a test that is severe by Popper's lights—even if it is highly *corroborated*—he regarded this as at most a report of the hypothesis' past performance and denied it afforded positive evidence for its correctness or reliability. In contrast, according to the severity principle, when a hypothesis H has passed a highly severe test we can infer that the data $\boldsymbol{x}$ provide good evidence for the correctness of H.

WHAT SHOULD THE ROLE OF FORMAL STATISTICAL IDEAS BE IN AN ACCOUNT OF EVIDENCE?

Chastened by the failures of purely formal, context-free, inductive logics, many philosophers of science nowadays tend to be critical of appealing to formal statistical ideas—Bayesian or frequentist (error statistical)—in erecting an account of evidence. Although we can agree with their doubts about purely formal E-R logics, failure to avail themselves of the methods and models of current statistical inference and modeling in science has been a major obstacle to developing philosophies of inference that are relevant to understanding and solving problems about evidence, both in science and in philosophy. In rejecting appeals to statistical methods as tantamount to advocating uniform, content-free approaches, philosophers of science have too readily given up on being able to say anything that is both general and relevant to the actual problems of evidence.

Are Statistical Methods Relevant Only to Formal Statistical Practice?

A related, and also mistaken, stance one often hears as grounds to reject, or at least minimize the relevance of, appeals to statistical methods in developing accounts of evidence and inference, is the supposition that such appeals could only be relevant for scientific inferences that explicitly make use of formal statistical ideas. But scientists evaluated evidence, the objection continues, before the development of statistical tools, and even now do not necessarily appeal to them (e.g., Chalmers, 2001). The flaw behind such objections is that they overlook the main philosophical goals behind appealing to statistical ideas; namely, to capture enough of the ingredients of scientific reasoning to solve philosophical problems about evidence and inference (Duhem's problem, underdetermination, theory-ladenness), and to understand and scrutinize various strategies and methodologies (e.g., preferring novel predictions, varying evidence, replication).

The Dean's Challenge

Peter Achinstein, a philosopher of science with whom I so often agree, likewise parts company with me when it comes to my advocating an appeal to statistics in building a philosophy of evidence. Achinstein concedes, in a pessimistic response to his dean's challenge as to the relevance of philosophy for science, that "standard philosophical theories about evidence are (and ought to be) ignored by scientists" (2001: 3). On the one hand, he is correct to declare philosophical accounts of evidence irrelevant to scientists, if those accounts view the question of whether data x provides evidence for H as a matter of purely logical computation. Whether data provide evidence for hypotheses, Achinstein rightly insists, is not an a priori but rather an empirical matter, and Achinstein, perhaps more than any other philosopher of science, has identified the inadequacies and counterexamples in existing E-R accounts.

On the other hand, one may reach a very different position from him as to what follows from such failures, and here is where we differ. He appears to take those failures to show that appealing to statistical ideas cannot provide the basis for a successful philosophical account of evidence. If the question of evidence is an empirical scientific matter, then in Achinstein's view philosophers can best see their job as delineating the concepts of evidence that scientists seem to use, leaving it up to scientists to apply them. However enlightening such conceptual analysis may prove to be, I think it is a mistake to curtail the philosopher's job in this fashion, and this mistake rests on the erroneous assumption that statistical accounts of evidence are restricted to supplying purely formal probabilistic computations. On the contrary, statistical ideas and methods—as used in practice—provide just the right blend of empirical and formal tools to provide forward-looking methods for appraising evidence. They can and should be at the heart of philosophical discussions aimed at resolving controversies about evidence in science—although we can grant that they generally have not been.

High Posterior Probability versus High Severity

The shortcomings Achinstein finds in probabilistic accounts of evidence serve to highlight key weaknesses in familiar Bayesian approaches: An increase in the posterior probability of H given x does not suffice for x to be evidence for H, even if coupled with his requirement that the posterior, $P(H/x)$, be high. However, by relying on such Bayesian assignments to locate the role for probability, Achinstein concludes—prematurely, in my opinion—that all statistical accounts fail. In particular, while he takes a necessary condition for x to be evidence for H that $P(H/x)$ be high, for some "objective" notion of probability, he requires also that there be the right kind of explanatory connection between x and H—something that he regards as going beyond any statistical assessments.

But why suppose that the work statistics can do for us is limited to such probabilistic computations? Rather than reason, as Achinstein seems to, that since the high posterior probability requirement fails to provide a sufficient condition for evidence, it follows that statistics cannot supply an adequate account of evidence, one may instead reject this requirement (which, after all, does not even enter into the assessment of error probabilities) and appeal to statistics to capture the severity requirement for evidence. That is the path I follow.

This appeal to statistics most nearly finds its home in non-Bayesian or frequentist accounts of statistics encompassing significance tests, (Neyman-Pearson) hypothesis tests, and estimation methods, as well as newer additions to this group of (error statistical) tools. In these accounts we find statistical methods that lend themselves to a conception of tests wherein probability arises, not to quantify support or probability in hypotheses, but to assess the probativeness (or reliability, or "trustworthiness") of the overall test procedure. From such accounts we learn at once that a methodology for severe testing cannot begin with "given" statements of evidence, but requires enough information about how the data were generated, and about the specific testing context, to assess the overall severity with which a claim or hypothesis may be inferred.

Warranting the Data

Given that "statistics claims to deal in a broad way with the collection and analysis of data" (Cox, 1981: 289), it should not be surprising that it would contain clues for a philosophical account of evidence. In particular, by viewing severity as the goal, we begin to see how statistical methods of data collection and analysis may be appealed to in order to get around what philosophers have typically considered overwhelming obstacles to justifying ampliative inference. While it is true that intermediary inferences are often required to arrive at inductive evidence, far from posing a threat to reliability, as is often thought, they may become the source of avoiding these very threats. Statistical ideas teach us how, by appropriately combining data, we may arrive at highly reliable claims from highly shaky data. The statistical cases encapsulate the reasoning in more qualitative contexts, as when the reports on the Columbia crash built inference upon inference to arrive at a clear pinpointing of blame.

A Severity (Re)interpretation of Neyman-Pearson (and Related) Methods

While retaining the central feature of these accounts, the more general "error-statistical" rubric frees us to reinterpret standard statistical accounts so as to avoid ever-present criticisms and misuses. In particular, we can reject the view of statistical hypotheses tests (e.g., Neyman-Pearson N-P tests) as mechanical tools with low long-run error probabilities, and construe them instead as tools for obtaining reliable experimental knowledge and severe tests.[12] The appeal to statistics in a philosophy of evidence, as I see it, is a two-way street: it gives insights into philosophical problems and confusions, while at the same time it serves to avoid fallacies and resolve debates in statistics. How to reinterpret N-P tests so that they supply post-data severity assessments that are sensitive to the actual outcome (unlike standard type I and type II error probabilities), and how this avoids recalcitrant problems, is discussed in detail elsewhere (Mayo, 1983, 1985, 1996, 2002b, c; Mayo and Spanos, 2000).

My concern in the remainder of this chapter is to reply to variations on a single type of criticism that has dogged non-Bayesian accounts, both in philosophy and statistical practice, namely, *that error statistical tests do not give us what we really want from an account of evidence, because they may regard* **x** *as good evidence for H, even though H is not accorded a high posterior probability*, according to one or another recommended way to obtain the requisite priors. This more general criticism can be and has been used as ammunition for a criticism directed specifically at the severity requirement, namely, that a hypothesis may have passed a severe test (it may be highly probed) even though it is not accorded high probability. The challenge revolves around what I have dubbed the "*highly probed versus highly probable debate.*"

After clarifying the notion of severity I have in mind, I argue, with respect to each variant of the criticism, that (1) the probabilistic assignment commits a fallacy, which may be called the *fallacy of instantiating probabilities*, and (2) the error statistical assessment of the data, but not the assessment advocated by the critic, is in sync with the goals of severity, and with our intuitions about when data should count as supporting evidence in science.

Achinstein's Criticism of the Severity Account of Evidence

I will begin with a simple variant on this criticism, as articulated by Peter Achinstein, and then develop and strengthen his charge in order to respond to the strongest Bayesian criticism in recent statistical literature.

We can get to the heart of the problem in short order: Achinstein's criticism assumes that (an appropriately random) sample resulting in 40% A's being B's suffices to pass, with severity, the statistical hypothesis that the population proportion of A's that are B's is .4—but this is false, at least in his example. Perhaps he is assuming that severity in the error statistical account is captured by what we termed the "simple comparative likelihood account," and admittedly, we said, Popper's view is open to such a reading. But this overlooks, and is at odds with, the central tenet of the severity account: *good fits (whether absolute or comparative) alone do not suffice for good tests!* Examining Achinstein's example and criticism serve to both illustrate the error statistical approach and set the stage for identifying key flaws in a whole cluster of Bayesian criticisms that run to this type.

Binomial Test T_1

Let us call the test he describes test T_1. A random sample of size n = 100 is taken, $\mathbf{X} = (X_1, \ldots, X_n)$, where each X_i is distributed as a Bernoulli random variable with unknown mean p, the probability of success, where in this case "success" means drawing a white ball in a random selection (he assumes that p is constant and trials are independent). We are to test two *simple* or *point* statistical hypotheses:

$$H_0\text{: p} = .4 \qquad \text{vs.} \qquad H_1\text{: p} = .6$$

Test T_1 may be described as a "point against point" test, since H_0 and H_1 each asserts just one of the possible parameter values. (H_0 is often called the "null" hypothesis, though none of my points turn on which one we regard as the null.) Such point vs. point (or simple against simple) tests are highly artificial and, strictly speaking, are not proper Neyman-Pearson tests because the hypotheses do not exhaust the parameter space, which includes all values from 0 to 1. However, because this is automatically taken into account in applying the severity criterion, we can proceed to the problem that Achinstein claims T_1 poses for my account.

Basic Concepts: Test Statistics, Significance Levels, Tail Areas

A statistical test is defined in terms of a *test statistic* or *distance measure* $d(\mathbf{X})$. In this example:

$$d(X) = (\bar{X} - p)/\sigma_x,$$

where $\bar{X}$ is the sample mean, and the sample standard deviation

$$\sigma_x = \sqrt{\frac{p(1-p)}{n}}$$

is about .05.[13] The outcome is given as $\bar{X}_{obs} = .4$.

To get the *significance level* of the observed difference, we ask: "How improbable is it to observe an $\bar{X}$ as far or farther from the value hypothesized in H_0, if in fact H_0 is true?"

To answer this we must calculate $P(\bar{X} > .4;\ p = .4) = P(d(X) > 0) = .5$. Since .5 is not small, this would yield a *non*statistically significant difference, so the test does *not* reject H_0. (Typically, it would be required that the statistical significance reach values as small as .05, or .01, corresponding to observing at least 50% or 55% white balls in this test.)

Note that calculating the significance level requires calculating not just the probability of the data point (.4) under H_0, but also the "tail area"—that is, the probability of outcomes *beyond* .4— under H_0. (It makes little difference in this test whether we consider > or ≥.)

The Criticism: First Variant

According to Achinstein, "On [Mayo's] view, . . . the result . . . (40 of the 100 balls selected are white) is good evidence for the hypothesis" H_0 (Achinstein, 2001: 134), because he supposes that I would regard H_0 as surviving a severe test. But has H_0 passed a severe test with outcome $\bar{X}_{\text{obs}} = .4$? No!

Let us abbreviate the severity with which hypothesis H passes test T_1 with outcome $\bar{X}_{\text{obs}}$ as:

$$SEV(H, \bar{X}_{\text{obs}}, T_1).$$

Although we can allow that $\bar{X}_{\text{obs}} = .4$ *fits* H_0, to calculate $SEV(H_0, \bar{X} = .4, T_1)$ requires calculating P(a "worse fit" with H_0; H_0 is false) ~ .5—and this is clearly not a high severity value! If one were to take this outcome as grounds for accepting H_0 we would erroneously do so 50% of the time! [*Note*: Since this is a non-significant result, the severity assessment happens to equal the observed significance level (or P-value), that is, .5.] To put this in other words, we may agree that H_0 would not be rejected by this outcome—but this is not tantamount to finding evidence *for* H_0. Indeed, taking no evidence against the null as evidence for it is a well-known fallacy.

To explain why, note that "H_0 is false" is the disjunction of values of p other than .4. Since the alternative statistical hypothesis in Achinstein's example, H_1, is in the positive direction, "H_0 is false" would generally be regarded as the one-sided alternative, $p > .4$. (The same argument can be made out if it is a two-sided alternative.) We cannot regard a failure to reject a null, that $p = .4$, as grounds for H_0: $p = .4$—that there is 0 discrepancy from .4—because the test would very probably have failed to reject H_0, even if in fact there *are* discrepancies from .4.

Suppose, for example, that the true proportion, $p = .41$. Of course we do not know the true value of p, but we need to consider the properties of our test under such hypothetical scenarios in order to ascertain what has and has not passed a test with severity. How severely can we say the data have ruled out a discrepancy of .01? In other words, What's the severity of the test that the hypothesis $p < .41$ may be said to have passed? We calculate

$$SEV(p < .41, \bar{X}_{\text{obs}} = .4, T_1) = P(\bar{X} > .4; p = .41) = P(D > -.2) \sim .6$$

But .6 is not very severe—40% of the time the test would fail to reject H_0 even if the underlying true value of p were .41. (We would typically want a severity of at least .9 before counting it as severe, although rather than select a single cut-off point, one may just report the severity obtained.) So we surely cannot say we have ruled out the existence of *any* positive discrepancy from .4, but without this we cannot say that the null hypothesis H_0 has passed a severe test. Therefore, we cannot regard such a result as good evidence for .4.[14]

However, we can find discrepancies that *are* ruled out with severity, and it is informative to calculate several, or even the entire severity curve.[15] Here are just a few:

$$SEV(\mathrm{p} < .5, \bar{X} = .4, T_1) \sim .97$$

$$SEV(\mathrm{p} < .6, \bar{X} = .4, T_1) \sim .999$$

So we *would* be entitled to rule out with severity that the true value of p was as great as .6, that is, infer $\mathrm{p} < .6$ (so $\boldsymbol{x}$ is evidence that $\mathrm{p} < .6$ and also that $\mathrm{p} < .5$). But Achinstein's alternative hypothesis H_1 asserts $\mathrm{p} = .6$, so we seem to be agreeing that our result is evidence against H_1! It is, but this does not constitute evidence for the hypothesis H_0. Fallacies stemming from applying tests to non-exhaustive hypotheses are legion (Mayo and Spanos, 2004).

The Criticism: A Bayesian Variant

Having agreed that our result is evidence against H_1, that p is as large as .6, we can mount the kind of criticism that Achinstein wishes to consider. It goes like this: For any hypothesis H that an error statistician such as myself regards as having passed a severe test, we can imagine that not-H has a high enough prior probability so that $P(\text{not-}H/x)$ is high, and thus $P(H/x)$ is low. Therefore, I would be claiming that there is evidence for H even though the posterior probability for H is low. (*Note*: Here we use the conditional probability symbol "/", since Achinstein is mounting a Bayesian criticism.)

To help Achinstein's criticism along as much as possible, let us put the evidential claim that we concurred with in a positive form: We have evidence that H': $\mathrm{p} < .6$. H' is a familiar complex statistical hypothesis, in contrast to the simple point hypothesis; it includes a disjunction of values. Now for the frequentist, we said, the hypothesis H is an assertion about the data-generating procedure (G) that gave rise to the observed data. Any such hypothesis would be regarded as correct or incorrect, even though these assertions can, and generally would, be qualified in terms of how good an approximation we have evidence for, for example, by reporting a margin of error. Doing so, however, would *not* be to assign a probability to any particular statistical hypothesis H.

The entire error statistical approach is deliberately designed to avoid appealing to priors, which are nearly always unavailable or irrelevant for scientific contexts. But the critic will proceed by trying to identify a prior probability for hypothesis H that even a frequentist can, allegedly, condone.

A Prior That Even a Frequentist Can Love?

Achinstein, to his credit, is seeking an objective account, and like many others, he assumes that the way to get objective prior probabilities is through relative frequencies of certain sorts. There are two or three main gambits used to obtain priors that are allegedly "kosher for frequentists," and we will consider them in turn. The first, the one to which Achinstein appeals, requires us to consider the population or data-generating procedure G, about which the statistical hypotheses make assertions, as itself selected from a population of populations. It is assumed that: *if we randomly select a population (i.e., a bag) from a population of populations (population of bags), p% of which have hypothesis H true of them, then the frequentist prior probability of H is p.*

This assumption, although plausible-sounding, is fallacious. Suppose our high school student, Isaac, has been randomly selected from a population of high school seniors, 10% of whom are college-ready. The probability of randomly selecting a college-ready student is .1, but this does not make the probability of Isaac being college-ready equal to .1.

A Particular Numerical Criticism

To turn to Achinstein's example, we are to consider an urn of bags (or populations) such that for each such bag a hypothesis H either is or is not true of it. In particular, if the bag selected is one with 60% white balls, then the hypothesis H_1 is true of that bag. In other words, H_1 plays the role of a one-place predicate H_1___, and any particular bag, say b, either has H_1 or not (i.e., either H_1b or ~ H_1b is true). Achinstein's criticism assumes:

> (*) If p% of the bags have property H_1, then for any randomly selected bag, b, $P(H_1) = $ p, (i.e., $P(H_1$,b$) = $ p).

In particular, the bag of balls from which we drew our sample of 100, bag b, "itself is chosen at random from a very large set of bags" (Achinstein, 2001: 134), out of which 99,999 out of 100,000 of the bags have 60% white balls. He infers:

> (*) P(H_1) = .99999, i.e., P(bag b has 60% white balls) = .9999.

Then applying Bayes's theorem, we get

(1) $P(H_1/\bar{X}_{\text{obs}} = .4) = .996,$

and he may wish to say

(2) $P(\text{not-}H_1/\bar{X}_{\text{obs}} = .4) = .004.$

But I have allowed earlier that

(3) $\bar{X}_{\text{obs}} = .4$ is evidence that H': $p < .6$.

Therefore, Achinstein's critique continues, Mayo claims we have good evidence that $p < .6$, and so, good evidence that not-H_1, even though the posterior probability of not-H_1 is very low.

Now this is problematic only if two additional premises are true:

> A necessary condition for x to be evidence for H is that $P(H/x) =$ high.
>
> The prior probability assignment in (*) (upon which the posterior is based) is valid.

I would deny both of these premises. Whereas the first is an issue of philosophy of evidence, the second is based on the fallacy I call *the fallacy of probabilistic instantiation*.

The Fallacy of Probabilistic Instantiation

Since the fallacy is committed repeatedly in mounting these criticisms, let us draw it out a bit. It is just the sort of misstep that one would hope philosophers of probability and statistics would use their acumen to expose, rather than commit.

The basis for the prior probability assignment in (*) is this:

1. Hypothesis H is true of p% of the populations (bags) in this urn of populations U.
2. P(H is true of a randomly selected bag from an urn of bags U) = p.
3. The randomly selected bag that was drawn in test T_1 is b_1.

Therefore

(*) $P(H \text{ is true of } b_1) = p$.

But either H is true of b_1 or not—the probability in (*) is fallacious and results from an unsound instantiation. It may help to make the point using confidence intervals.

An Analogous Fallacy with Confidence Intervals

A 95% confidence interval estimation procedure has a probability of covering the true but unknown value of a parameter equal to .95, in a series of experiments on the same or different populations. Each bag in the pool of bags is a different population, and a 95% confidence interval estimate may be formed for each; however, each interval estimate either will or will not be true of that population. Nevertheless, the foregoing reasoning would countenance the fallacious inference:

1. P (the 95% confidence interval procedure yields an interval estimate that is true) =.95
2. The 95% confidence interval procedure yields an interval estimate: $(.3 < p < .5)$, let us suppose.

Therefore,

(*) $P(.3 < p < .5) = .95$.

But either $(.3 < p < .5)$ or not!

Students from the Wrong Side of Town

Examples of balls in bags, however dear to philosophers, are rather distant from the kinds of realistic examples about which one's intuitions are clearest. It will be useful to turn to some more realistic examples on which the identical criticism has been based, both against frequentist statistics in general and my severity account in particular. Once again, the fallacy of instantiating probabilities is committed.

Our student, Isaac, has passed comprehensive tests of mastery of high school subjects regarded as indicating college readiness. Because such high scores x could rarely result among high school students who are not sufficiently prepared to be deemed college-ready, we regarded x as good evidence for

> *H*: Isaac is not deficient but is college-ready.

Thus x is good evidence against

> *H*′: Isaac's mastery of high school subjects is deficient, that is, he is not college-ready.

Although as with the binomial example, we would ordinarily consider degrees of readiness, we can keep to this oversimplified rendition to go along with our critic as much as possible.

In this variation of the Bayesian example (from Howson, 1997 and others), we are given that

> $P(x/H$: Isaac is college-ready) is practically 1

whereas

> $P(x/H'$: not college-ready (i.e., deficient)) = very low, say .05.

Given the assumptions that go into the probability calculations are met, it would seem that H has stood up to a fairly severe test.

"But wait a minute!" says the critic. Isaac was randomly selected from a population (perhaps a certain section of the Bronx) wherein college readiness is exceedingly rare—say, only .1% would be correctly described as ready. Accordingly, it is reasoned that

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

Thus, the posterior probability for *H* is still low, and for the alternative, H′ (deficient), the posterior is high. In particular, to give one illustration, we can have:

> $P(H'/x) = .95$.

Although $P(H/x)$ has increased from $P(H)$, the posterior for *H* is still low because of the extremely small prior for *H*.

Notice that if Isaac had been selected from a population where college readiness is rather common—if, say, he was selected from students in a certain affluent neighborhood—then the very same set of passing scores x would now be regarded as strong evidence for *H*, Isaac being ready. Using this way of evaluating evidence, a high school student from a non-affluent neighborhood would need to have scored quite a bit higher on these tests than one selected from the affluent neighborhood in order for his scores to be considered evidence for his readiness! (Talk about reverse discrimination!)

Once again the same fallacy is committed in arriving at (*), only here perhaps our intuitions that something is amiss are more pronounced.

Although the probability of a random sample (i.e., a student) taken from the "urn" of highschoolers in the given area of the Bronx is .001, it does not follow that Isaac, the one we happened to select, has a probability of .001 of being college-ready (see Mayo, 1997).

Concluding the Achinstein Discussion

The foregoing remarks do not preclude saying that the probability of a student randomly selected (from a given population U where p% are ready) is ready equals p—so long as one is clear about the kind of trial is described. What they preclude is treating the probability of the occurrence of this *event* as if it provided a legitimate frequentist probability for the *hypothesis* H in the testing problem. The cavalier manner in which philosophers of probability use the term "hypothesis" makes it too easy to slip between events and statistical hypotheses. A statistical hypothesis, for a frequentist statistician, must describe enough about the data generation G to assign probabilities to all possible outcomes; an event does not.

Nor need we preclude the possibility of a statistical hypothesis H having a legitimate frequentist prior. For example, a frequentist probability that Isaac is college-ready might refer to genetic and environmental factors that determine the chance that a high school student (from a specified population) is deficient—something we can scarcely even cash out, much less compute. I return to these points in my concluding comments.

In fairness to Achinstein, it must be noted that he too, presumably, denies that the high posterior for H entails that x provides evidence for H, but not because he questions the prior probability assignment in (*). He accepts (*) but denies that high posterior probability suffices for evidence. Why then, one might ask, should it even be a necessary requirement; that is, why suppose the computation based on (*) is necessary for evidence for H? By assuming a high posterior is necessary, Achinstein must look elsewhere to avoid taking x as evidence for H in examples such as test T_1; and he does so by appealing to intuitions that tell him that the right sort of explanatory connection is absent in these examples. By leaving this at a vague intuitive level, however, his account supplies no directions for determining if one really has evidence. We are left to fall back on the very intuitions we wish an account of evidence to provide. Error probabilities, within a severity assessment, let us go further.

For instance, the severity criterion provides the basis for denying that the high posterior $P(H'/x)$ is evidence for H', in the case of Isaac's readiness (see "Students from the Wrong Side of Town," above). The error probability associated with such a procedure for interpreting data would be extremely high. Thus it would lead to regarding data as evidence on the basis of a procedure that very probably would be wrong—low severity! A severity assessment captures and guides intuitions about evidence.

What We *Really,* Really Want Is . . . High Severity!

It is (or should be) well known that error probabilistic concepts, such as p-values and type I and type II errors, do not supply probabilities to statistical hypotheses, and that interpreting them as if they did leads to fallacies, paradoxes, and contradictions. Error probabilities and any severity assessment that we would base upon them, are—quite deliberately—defined exclusively in terms of the *sampling distribution* of $d(\boldsymbol{X})$, under one or another statistical hypothesis of interest. In contrast, posterior probabilities of a hypothesis such as H_0, conditional on the observed $d(x)$ require a prior probability assignment to (an exhaustive set of) hypotheses. (The capital $\boldsymbol{X}$ indicates the random variable; the lower case x, the resulting value or outcome.) Nevertheless, critics—especially from Bayesian quarters—have long insisted that "what we really want" from tests are posterior probabilities of hypotheses, and some even argue that testers cannot help but fallaciously interpret p-values as supplying a posterior to the null. Those criticisms are analogous to those in "Achinstein's Criticism . . ." (above) and commit analogous fallacies.

A Common Variant on the Criticisms: P-Values versus Posterior Probabilities

The most telling criticisms are put in terms of p-values: Critics argue that (a) certain choices of prior probabilities for the null and alternative hypotheses show that a small p-value is consistent with a much higher posterior probability in a null hypothesis, from which they conclude that (b) significance test reasoning is invalid, or at least is incapable of being used to assess the evidence against the null hypothesis. The criticism assumes that the Bayesian posterior gives the correct or even an acceptable measure of the degree of evidence, reliability, or beliefworthiness properly accorded the null, and thus a conflict between Bayesian and frequentist assessments shows the latter to be at fault! Nowhere is this assumption defended—one is to have a gut feeling that the only way to use data to bear upon the truth of hypotheses is by means of a posterior probability assignment. If the Bayesian posteriors really did provide assessments of the reliability or beliefworthiness of hypotheses, that would be one thing—but they do not.

Achinstein at least is aware of the need to provide grounds for requiring, as necessary for x to be evidence for H, that $P(H/x)$ be high. His reasons are persuasive—*provided that "high probability" equates to "highly warranted in believing."* The trouble is, the prior and posterior probabilities he actually calculates do not warrant this equation. The same will be true for the criticisms from more formal quarters.

(Two-sided) Test of a Mean of a Normal Distribution, Test T_2

The most influential attempts to demonstrate the conflict between p-values and Bayesian posteriors consider a two-sided Normal distribution test, test T_2, of H_0: $\mu = 0$ versus H_1: $\mu \neq 0$ (the difference between p-values and posteriors being less marked with one-sided tests), as in Pratt (1965), Berger and Sellke (1987), and Berger (2003). A random sample $X = (X_1, \ldots, X_n)$ is taken where each X_i is distributed Normally with unknown mean μ, and known standard deviation σ and we are testing for discrepancies from 0 in both the positive and negative directions (see Mayo, 2003). We can imagine that null hypothesis H_0 is (the formal embodiment of) the claim:

> H_0: There are no increased risks (or benefits) associated with HRT in women treated for 10 years.

A familiar criticism even with sample size only 50 is: "If $n = 50$ one can classically 'reject H_0 at significance level $p = .05$,' although $P(H_0/x) = .52$ (which would actually indicate that the evidence favors H_0)" (Berger and Sellke, 1987: 113; we replace Pr with P for consistency) assuming a prior of $P(H_0) = .5$. Note that we would take the low significance level as evidence against H_0 and for H_1, evidence for the very weak claim that there is *some* non-zero difference in risk rates between treated and non-treated. This is taken as a criticism of p-values, only because it is assumed that the .51 posterior in H_0 is the evidence for H_0. (The severity interpretation in the case of a rejection automatically goes on to consider the extent of the discrepancy indicated, but we can put that aside for the present discussion.)

We can concede the critic's point that data we would regard as evidence against H_0 would, on the Bayesian construal, "actually indicate that the evidence favors H_0," at least assuming the recommended prior of .5 to H_0. As the sample size increases, the conflict becomes more noteworthy. If $n = 1000$, a result statistically significant at the .05 level leads to a posterior to the null of .82! It has again gone up, but even more dramatically. Nevertheless, far from discrediting the frequentist assessment, this fact seems to us to count against regarding the Bayesian posterior as properly measuring evidence. This leads to the question, What warrants the prior probability on which the example is based?

The Case of Subjective Priors

Many Bayesians construe prior probability assignments as quantifying their degrees of belief in hypotheses, extracted either by intuition or by strategies of eliciting betting behavior. A strong prior degree of belief in H_0 suffices to ensure that the posterior of H_0 is still high, even with a statistically significant result. By setting the prior high enough, not surprisingly, the 0-risk H_0 is saved from counterevidence. That is hardly to subject H_0 to a severe test. Under the pretense of giving a "fair" assignment of priors, assigning .5 to H_0 gives so much weight to H_0 that even highly significant observed departures are taken as strengthening the degree of belief in the null. If understood as a report of an agent's actual degrees of belief, then it is hard to fault, but if we want to know how strongly we *ought* to believe in a hypothesis on the basis of data, the subjective Bayesian analysis will not do.

The Assumption of "Objective" Bayesian Priors

Bayesians claim to have arrived at the chart in Table 6.1 and several related charts by appealing to one or another alleged "objective" priors. In particular, following Jeffreys (1939), a prior probability assignment of .5 is given to H_0 and the remaining .5 probability is spread out over the alternative parameter space. Having picked out the point null as a value of special interest to test, their reasoning goes, a "spiked concentration of belief in the null" is "impartial"—but is it? The frequentist says no. Moreover, the "spiked concentration of belief in the null" is at odds with the role played by null hypotheses in testing, where it is so often assumed that "all nulls are false," and we wish only to learn (via significance tests) "how false" the null is. Thus this spiked concentration of belief should hardly be taken as the guidepost from which to launch a critique of significance tests.

Table 6.1 $P(H_0/x)$ *for Jeffreys-Type Prior*

					n			
p	*t*	1	5	10	20	50	100	1,000
.10	1.645	.42	.44	.47	.56	.65	.72	.89
.05	1.960	.35	.33	.37	.42	.52	.60	.82
.01	2.576	.21	.13	.14	.16	.22	.27	.53
.001	3.291	.086	.026	.024	.026	.034	.045	.124

Source: Berger and Sellke (1987: 113). Used by permission.

Even a discrepancy of several standard deviations away from 0, as Table 6.1 shows, will have little or no chance of being considered as *some* evidence against the null because the posterior probability is below the prior! The error statistician would criticize such tests as having extremely low probability of detecting a false null, that is, low power. Such tests would commit type II errors very often if not in extreme cases with probability 1.[16] This is a remarkable procedure. In the HRT study, there were more than 16,000 women, so the discrepancy between the p-value and posteriors would be even more marked than with n = 1000.

One can see why the Bayesian significance tester wishes to start with a fairly high prior to the null—else, a rejection of the null 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, even assuming it is not precisely true, because we can learn "how false" it is. It is all well and good to exhort that what we really want is an assessment of the truth of H_0 given the data—but the Bayesian posterior hardly seems appropriate for the task. A dialogue between a researcher and a Bayesian might proceed as follows:

> RESEARCHER: I have found a difference that is significant at the .01 level. Since this would occur only 1% of the time, if in fact there was no effect, it seems to indicate evidence of a genuine effect.
>
> BAYESIAN: This just goes to show what's wrong with significance test reasoning. For a Bayesian who assigned a "fair" prior to the null, one's degree of belief in H_0 would go from .5 to .82.
>
> RESEARCHER: But the null hypothesis would be outside of the 95% confidence interval—surely that is evidence against it, and yet you assign it a probability of .82.
>
> BAYESIAN: Only by calculating a Bayes factor (or related conditional measure) can one judge how well the data supports a hypothesis (Berger and Delempady). The Bayesian analysis tells you that the posterior probability of H_0 is .82—and isn't that what you really want to know?
>
> RESEARCHER: Hmm—I don't think so . . .

"Frequentist" Priors Which Are Not Kosher for Frequentist Error Statisticians

A common gambit by Bayesian critics is to construct examples where allegedly only frequentist priors are used. Again considering a Normal distribution test of H_0: $\mu = 0$ versus H_1: $\mu \neq 0$, Bayesians assure us that even one who insists on a frequentist interpretation of probabilities can see that low p-values may be overstating the evidence against the null. In particular, to construe a hypothesis as a random variable, it is imagined that we sample randomly from a population of hypotheses, some proportion of which are assumed to be true. This is a variation on the technique for erecting a frequentist prior that we saw in "Achinstein's Criticism . . . ," and the same fallacy of instantiating probabilities is committed.

We are to consider a pool of null hypotheses from which H_0 may be seen to belong, and compute the proportion of these that have been found to be true in the past. 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, and 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? Shall we look at all those asserting no increased risk or benefit of any sort? Or no increased risk of specific diseases, such as clotting disorders or breast cancer? In men and women? Or in women only? With hormonal drugs, or any treatments? The percentages "initially true" 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 have thus far not been rejected by other statistical tests! (See Mayo, 2003.)

Innocence by Association

Further, even if we agreed that there was a 50% chance of randomly selecting a true null hypothesis from a given pool of nulls, that would still not give the error statistician a frequentist prior probability of the truth of *this* hypothesis, for example, that HRT has no effect on breast cancer risks. Either HRT alters cancer risk or it doesn't. The relevant parameter—say, the increased risk of breast cancer—could conceivably be modeled as a random variable, but its distribution would not be given by computing the rates of other apparently benign or useless treatments! This "frequentist" Bayesian analysis assumes a kind of "innocence by association," wherein a given H_0 gets the benefit of having been drawn from a pool of true or not-yet-rejected nulls. Perhaps the tests have been insufficiently sensitive to detect risks of interest. Why should that be grounds for denying there is evidence of a genuine risk with respect to a treatment (e.g., HRT) that *does* show statistically significant risks?

What Is Varying and What Is Held Fixed

For the error statistician, it must be remembered, what varies is $d(x)$—what is fixed is the particular hypothesis of interest. By contrast, the

Bayesian analysis is conditional on a value of $\boldsymbol{x}$ (other values that could have resulted but did not are irrelevant once $\boldsymbol{x}$ is in hand);[17] $\boldsymbol{X}$ is fixed and the hypotheses are taken to vary.[18] We need not deny that there are contexts wherein that kind of calculation is relevant, but no case has been made that evaluating the truth or correctness of *this* hypothesis is one of those contexts.

Fisher: The Function of the P-Value Is Not Capable of Finding Expression

Faced with conflicts between error probabilities and Bayesian posterior probabilities, the error probabilist would conclude that the flaw lies with the latter measure. This is precisely what Fisher argued, and it seems fitting to finish our story by considering him.

Discussing a test of the hypothesis that the stars are distributed at random, Fisher takes the low p-value (about 1 in 33,000) to "exclude at a high level of significance any theory involving a random distribution" (Fisher, 1956: 42). Even if one were to imagine that H_0 had an extremely high prior probability, Fisher continues—never mind "what such a statement of probability a priori could possibly mean"—the resulting high posteriori probability to H_0 would only show, he thinks, that "reluctance to accept a hypothesis strongly contradicted by a test of significance" (ibid: 44) . . . "is not capable of finding expression in any calculation of probability a posteriori" (ibid: 43). Note too that frequentists do not deny there is ever a legitimate frequentist prior probability distribution for a statistical hypothesis. One may consider hypotheses about such distributions and subject them to probative tests. Indeed, if one were to consider the claim about the a priori probability to be itself a hypothesis, Fisher suggests, it would be rejected by the data!

Concluding Comments

In this chapter I have attempted to sketch a conception of evidence as the result of surviving a severe or risky test:

> Data x in test T provide good evidence for inferring H (just) to the extent that hypothesis H has passed a severe test T with x.

On this conception of evidence, even claims that are highly beliefworthy, or highly probable (however probability is construed), would not get credit from tests that poorly probed them. Evidential credit goes only to those specific hypotheses or models that have survived probative testing. I showed how this notion should be cashed out, and contrasted it to other severity accounts, for example, the comparative account of Popper. I then confronted central criticisms raised by philosophers such as Achinstein and Howson, as well as by statistical practitioners such as Berger, Cohen, or Meehl, regarding the error statistical account of tests. Whether the criticisms are directly aimed at severity or, as in statistics, at error probabilistic tests, the criticisms, if sound, would show the inadequacy of the severity account of evidence that I favor. I grant that hypotheses that pass severe tests on my account may be accorded low posterior degrees of support, probability, or belief, while denying this shows any inadequacy in my account.

I have argued, with respect to each variant of the criticism, (1) that the probabilistic assignment commits a fallacy, which may be called the *fallacy of instantiating probabilities*; and (2) that the error statistical assessment of the data, but not the assessment advocated by the critic, is in sync with the goals of severity, and with our intuitions about when data should count as supporting evidence in science. *Highly probed* differs from *highly probable* (in any of the interpretations put forward for the latter), ***and it is the former that matters for evidence.***

NOTES

[1] Robert A. Wilson, M.D., *Feminine Forever* (M. Evans and Company Inc., New York, 1966).

[2] AARP, American Association of Retired Persons, Nov/Dec 2002, p. 72.

[3] Minimally $P(x;H) > P(x;\text{not-}H)$. The advantage of this definition is that any measure of evidential relationship, degree of confirmation, probability, etc., can be regarded as supplying a fit measure. Severity can then be assessed by computing the error probability required in (ii).

[4] Ben-Haim makes this notion rigorous by means of a definition based on convex sets, but it is one that I do not understand sufficiently to explicate.

[5] For example, searching for factors that show statistically significant correlations increases the overall error rate, in contrast to accounts that endorse the likelihood principle. See Mayo and Kruse, 2001; Cox and Hinkley, 1974.

[6] This contrasts with the use of type I and type II error probabilities in Neyman-Pearson tests: low type I error is desirable when the null hypothesis is rejected; low type II error, when it is accepted.

[7] Of course the severity analysis does not in itself specify the type or size of discrepancies that are of substantive importance, nor what should be done when discrepancies are found. However, this portion of the analysis sets the stage for subsequent strategies for respecifying models to arrive at empirically adequate statistical models (Mayo and Spanos, 2004).

[8] This is to be cashed out as "the probability of x under the assumption that H incorrectly describes the procedure that actually generated data x."

[9] There are qualifications here that I am omitting.

[10] The idea that "same likelihood" means "same evidence" is the thrust of the likelihood principle that underlies "Likelihoodist" and Bayesian accounts (for a full discussion, see Mayo and Kruse, 2001; Mayo, 2002b).

[11] Had Popper been aware of the developments in statistics going on right around him in the 1930s and 40s, one suspects the history of the philosophy of science would have been very different. In a letter from Popper in the early 1990s, he expressed regret at having never fully studied statistical methodology.

12. How this contrasts with their traditional construal as mechanical tools for controlling error-rates in the long run is a complex issue that is discussed at length elsewhere (see Mayo, 1996).

13. It must be calculated under one of the two hypotheses. Under H_0 it is .05; under H_1, slightly less.

14. That is the purpose of the "rule of acceptance" in the severity interpretation of statistical tests (no evidence against is not evidence for). See Mayo, 1996.

15. See Mayo and Spanos, 2000.

16. For example, consider a test with significance level .05. Such a test finds evidence for H_0 if the result does not reject H_0, and still finds evidence for H_0 when the usual test rejects H_0 at the .05 level-hence, it has no chance of finding evidence against H_0. Such a test would have 0 severity if used to infer H_0.

17. Such irrelevance of the sample space follows from the Likelihood Principle.

18. Admittedly, some frequentist philosophers (Salmon, following Reichenbach) pursued the idea of giving frequentist priors this way; though Reichenbach regarded the ability to make "claims about how often hypotheses like this one are true" as at most a possible situation we might find ourselves in once we had learned enough. Even if we had such numbers, their relevance to reasoning about the hypothesis in hand would be dubious (Mayo, 1996, ch. 4).

REFERENCES

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

Ben-Haim, Y. (2001) *Information-Gap Decision Theory: Decisions Under Severe Uncertainty*, Academic Press, San Diego.

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

Berger, J. O., and T. Sellke (1987) "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence," *Journal of the American Statistical Association*, 82: 112–22.

Birnbaum, A. (1962) "On the Foundations of Statistical Inference" (with discussion), *Journal of the American Statistical Association*, 57: 269–326.

Carnap, R. (1962) *Logical Foundations of Probability*, University of Chicago Press, Chicago.

Casella, G., and R. L. Berger (1987) "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," *Journal of the American Statistical Association*, 82: 106–11.

Chalmers, A. F. (1999) *What Is This Thing Called Science?* 3rd ed., University of Queensland Press, Australia.

Cohen, J. (1994) "The Earth Is Round (p < .05)," *American Psychologist* 49 (12): 997–1003.

Cox, D. R. (1981) "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14: 325–31.

Cox, D. R., and D. V. Hinkley (1974) *Theoretical Statistics*, Chapman and Hall, London.

Dorling, J. (1979) "Bayesian Personalism, the Methodology of Scientific Research Programmes, and Duhem's Problem," *Studies in History and Philosophy of Science*, 10: 177–87.

Earman, J. (1992) *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, MIT Press, Cambridge, MA.

Edwards, W., H. Lindman, and L. Savage (1963) "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70: 193–242.

Efron, B. (1986) "Why Isn't Everyone a Bayesian?" *The American Statistician*, 40:1–4.

Fisher, R. A. (1930) "Inverse Probability," *Proceedings of the Cambridge Philosophical Society*, 26: 528–35.

10

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

Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.

Gibbons, J. D. and J. W. Pratt (1975) "P-values: Interpretation and Methodology," *The American Statistician*, 29: 20–25.

Hacking, I. (1965) *Logic of Statistical Inference*, Cambridge (CVP).

Hacking, I. (1980) "The Theory of Probable Inference: Neyman, Peirce and Braithwaite," pp. 141–60 in D. H. Mellor, ed., *Science, Belief and Behavior: Essays in Honour of R.B. Braithwaite*, Cambridge University Press, Cambridge.

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

Howson, C. (1997) *Philosophy of Science*, 64: 268–90.

Howson, C. and P. Urbach (1989) *Scientific Reasoning: The Bayesian Approach*, Open Court. La Salle, IL (Second Edition, 1993).

Kyburg, H. E., Jr. (1993) "The Scope of Bayesian Reasoning," pp. 139–52 in D. Hull, M. Forbes, and K. Okruhlik, eds. *PSA 1992*, Vol. II, Philosophy of Science Association, East Lansing, MI.

Kyburg, H. E., Jr., and M. Thalos, eds. (2002) *Probability Is the Very Guide of Life*, Open Court, Oxford.

Laudan, L. (1997) "How About Bust? Factoring Explanatory Power Back into Theory Evaluation," *Philosophy of Science* 64: 303–16.

Lehmann, E. L. (1990) "Model Specification: the views of Fisher and Neyman, and later developments," *Statistical Science*, 5: 160–68.

Lehmann, E. L. (1993) "The Fisher and Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, 88: 1242–49.

Lindley, D. V. (1957) "A Statistical Paradox," *Biometrika*, 44: 187–92.

Lindley, D. V. (1976) "Bayesian Statistics," in W. L. Harper and C. A. Hooker, eds. (1976), pp. 353–62.

Mayo, D. G. (1983) "An Objective Theory of Statistical Testing," *Synthese*, 57: 297–340.

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

Mayo, D. G. (1996) *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

Mayo, D. G. (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*, 64: 222–24 and 323–33.

Mayo, D. G. (2002a) "Theory Testing, Statistical Methodology, and the Growth of Experimental Knowledge," pp. 171-90 in *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science*, Kluwer Press, Netherlands.

Mayo, D. G. (2002b) "An Error-Statistical Philosophy of Evidence" and "Response to Professors McCoy and Casella," pp. 79–97, 101–115, in *Scientific Evidence*, Contributions to the Ecological Society of America Conference, University of Chicago Press.

Mayo, D. G. (2002c) "Severe Testing as a Guide for Inductive Learning," pp. 89–117 in H. E. Kyburg and M. Thalos, eds., *Probability Is the Very Guide of Life* (2002), Open Court, Chicago.

Mayo, D. G. (2003) "Could Fisher, Jeffreys, and Neyman Have Agreed? Commentary on J. Berger's Fisher Address," *Statistical Science* 18: 19–24.

Mayo, D. G. (2004) *The Philosophy of Statistics*, Routledge Encyclopedia.

Mayo, D. G., and A. Spanos (2000) "A Post-data Interpretation of Neyman-Pearson Methods Based on a Conception of Severe Testing," *Measurements in Physics and Economics Discussion Paper Series*, History and Methodology of Economics Group, The London School of Economics and Political Science, Tymes Court, London.

Mayo, D. G., and A. Spanos (2004) "Methodology in Practice: Statistical Misspecification Practice," *Philosophy of Science* 71 (5).

Mayo, D. G., and M. Kruse (2001) "Principles of Inference and their Consequences," pp. 381–403 in D. Cornfield and J. Williamson, eds., *Foundations of Bayesianism*, Kluwer Academic Publishers, Netherlands.

Meehl, P. E. (1967/1970) "Theory-Testing in Psychology and Physics: A Methodological Paradox," In D. E. Morrison and R. E. Henkel, eds., *The Significance Test Controversy* (1970), Aldine, Chicago.

Morrison, D., and R. Henkel, eds. (1970) *The Significance Test Controversy*, Aldine, Chicago.

NAS Report (Columbia Accident Investigation Board Report, August 2003, www.caib.us/news/report/default.html).

Neyman, J. (1935) "On the Problem of Confidence Intervals," *The Annals of Mathematical Statistics*, 6: 111–16.

Neyman, J. (1941) "Fiducial Argument and the Theory of Confidence Intervals," *Biometrika* 32: 128–50.

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

Neyman, J. (1955) "The Problem of Inductive Inference," *Communications on Pure and Applied Mathematics*: 13–46.

Neyman, J. (1957a) "Inductive Behavior as a Basic Concept of Philosophy of Science," 25: 7–22.

Neyman, J. (1976) "Tests of Statistical Hypotheses and their use in Studies of Natural Phenomena," *Communications in Statistics—Theory and Methods*, 5: 737–51.

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

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

Pearson, E. S. (1955) "Statistical Concepts in Their Relation to Reality," *Journal of the Royal Statistical Society*, B, 17: 204–07.

Peirce, C. S. (1931–35) *Collected Papers*, Vols. I-VI, ed. C. Hartshorne and P. Weiss; Vols. VII-VIII, ed. A. Burks, Harvard University Press, Cambridge, MA.

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

Popper, K. (1983) *Realism and the Aim of Science*, Rowman and Littlefield, Totawa, NJ.

Pratt, J. (1965) "Bayesian Interpretation of Standard Inference Statements" (with discussion), *Journal of the Royal Statistical Society*, B, 27: 169–203.

Rosenthal, R. (1994) "Parametric Measures of Effect Sizes," pp. 231–44 in H. M. Cooper and L. V. Hedges, eds., *The Handbook of Research Synthesis*, Sage, Newbury, CA.

Rosenthal, R., and J. Gaito (1963) "The Interpretation of Levels of Significance by Psychological Researchers," *Journal of Psychology*, 64: 725–39.

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

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

Savage, L., ed. (1962) *The Foundations of Statistical Inference: A Discussion*, Methuen, London.