CHAPTER SIX

Severe Tests and Methodological Underdetermination

> The basic trouble with the hypothetico-deductive inference is that it always leaves us with an embarrassing superabundance of hypotheses. All of these hypotheses are equally adequate to the available data from the standpoint of the pure hypothetico-deductive framework.
>
> —W. Salmon, *The Foundations of Scientific Inference*, p. 115

A MAJOR PROBLEM that has been thought to stand in the way of an adequate account of hypothesis appraisal may be termed the *alternative hypothesis objection:* that whatever rule is specified for positively appraising *H*, there will always be rival hypotheses that satisfy the rule equally well. Evidence in accordance with hypothesis *H* cannot really count in favor of *H*, it is objected, if it counts equally well for any number of (perhaps infinitely many) other hypotheses that would also accord with *H*.

This problem is a version of the general problem of underdetermination of hypotheses by data: if data cannot unequivocally pick out hypothesis *H* over alternatives, then the hypotheses are underdetermined by evidence. Some have considered this problem so intractable as to render hopeless any attempt to erect a methodology of appraisal. No such conclusion is warranted, however. There is no general argument showing that all rules of appraisal are subject to this objection: at most it has been successfully waged against certain specific rules (e.g., the straight rule, simple hypothetico-deductivism, falsificationist accounts). Since chapter 1 I have been hinting that I would propose utilizing a test's severity to answer the underdetermination challenge. It is time to make good on this promise. Doing so demands that we be much clearer and more rigorous about our notion of severity than we have been thus far. Indeed, by exploring how an account of severe testing answers the alternative hypothesis objection, we will at the same time be piecing together the elements needed for understanding the severity notion. In anticipation of some of my theses, I will argue that

1. the existence of hypotheses alternative to *H* that entail or accord with evidence *e* (as well as or even better than *H*) does not prevent *H* from passing a severe test with *e;*
2. computing a test's severity does not call for assigning probabilities to hypotheses;
3. even allowing that there are always alternative hypotheses that entail or fit evidence *e*, there are not always alternatives equally severely tested by *e*.

As important as many philosophers of science regard the alternative hypothesis challenge, others dismiss it as merely a " philosopher's problem," not a genuine problem confronting scientists. In the latters' view, scientists strive to find a single hypothesis that accounts for all the data on a given problem and are untroubled by the possibility of alternatives. Granted, there are many examples in which it is generally agreed that any alternative to a well-tested hypothesis *H* is either obviously wrong or insignificantly different from *H*, but this enviable situation arises only after much of the work of ruling out alternatives has been accomplished. Anyone seeking an account adequate to the task of *building up* experimental knowledge, as I am, must be prepared to deal with far more equivocal situations. Moreover, an adequate philosophical account should be able to explain how scientists are warranted, when they are, in affirming one hypothesis over others that might also fit the data.

Grappling with the alternative hypothesis objection will bear other fruit. Appealing to a test's severity lets us see our way clear around common misinterpretations of standard statistical tests. In section 6.5, for example, the question of how to interpret statistically insignificant differences is addressed.

6.1 METHODOLOGICAL UNDERDETERMINATION

The "alternative hypothesis objection" that concerns me needs to be distinguished from some of the more radical variants of underdetermination. Some of these more radical variants are the focus of a paper by Larry Laudan (1990a), "Demystifying Underdetermination." "[O]n the strength of one or another variant of the thesis of underdetermination," Laudan remarks, "a motley coalition of philosophers and sociologists has drawn some dire morals for the epistemological enterprise." Several examples follow.

> Quine has claimed that theories are so radically underdetermined by the data that a scientist can, if he wishes, hold on to *any* theory he likes, "come what may." Lakatos and Feyerabend have taken the un-

> derdetermination of theories to justify the claim that the only difference between empirically successful and empirically unsuccessful theories lay in the talents and resources of their respective advocates. . . . Hesse and Bloor have claimed that underdetermination shows the *necessity* for bringing noncognitive, social factors into play in explaining the theory choices of scientists. (Laudan 1990a, p. 268)

Laudan argues that the Quinean thesis that "any hypothesis can rationally be held come what may" as well as other strong relativist positions are committed to what he calls the *egalitarian thesis.* "It insists that: *every [hypothesis] is as well supported by the evidence as any of its rivals*" (p. 271). Nevertheless, Laudan maintains that a close look at underdetermination arguments shows that they at most sustain a weaker form of underdetermination, which he calls *the nonuniqueness thesis.* "It holds that: *for any [hypothesis* H*] and any given body of evidence supporting* [H], *there is at least one rival* (*i.e., contrary*) *to* [H] *that is as well supported as* [H]" (p. 271). Laudan denies that the nonuniqueness thesis has particularly dire consequences for methodology; his concern is only with the extreme challenge "that the project of developing a methodology of science is a waste of time since, no matter what rules of evidence we eventually produce, those rules will do nothing to delimit choice" (p. 281). I agree that the nonuniqueness thesis will not sustain the radical critique of methodology as utterly "toothless," but I am concerned to show that methodology has a severe bite!

Even if it is granted that empirical evidence serves *some* role in delimiting hypotheses and theories, the version of underdetermination that still has to be grappled with is the alternative hypothesis objection with which I began, that for any hypothesis *H* and any evidence, there will always be a rival hypothesis equally successful as *H*. The objection, it should be clear, is that criteria of success based on methodology and evidence alone underdetermine choice. It may be stated more explicitly as the thesis of methodological underdetermination (MUD):

> *Methodological underdetermination:* any evidence taken as a good test of (or good support for) hypothesis *H* would (on that account of testing or support) be taken as an equally good test of (or equally good support for) some rival to *H*.

While not alleging that anything goes, it is a mistake to suppose that the MUD thesis poses no serious threat to the methodological enterprise. The reason formal accounts of testing and confirmation ran into trouble was not that they failed to delimit choice at all, but that they could not delimit choice sufficiently well (e.g., Goodman's riddle). Moreover, if hypothesis appraisal is not determined by methodology and evidence, then when there is agreement in science, it would seem to be the result of extraevidential factors (as Kuhn and others argue).

The existence of alternative hypotheses equally well tested by evidence need not always be problematic. For example, it is unlikely to be problematic that a hypothesis about a continuous parameter is about as well tested as another hypothesis that differs by only a tiny fraction. In the following discussion of my account of severe testing, I will focus on the seemingly most threatening variants of the MUD challenge.

Clearly, not just any rule of appraisal that selects a unique hypothesis constitutes an adequate answer to the challenge. Not just any sort of rule is going to free us from many of the most troubling implications of MUD. That is why the Bayesian Way does not help with my problem. Its way of differentially supporting two hypotheses that equally well entail (or otherwise fit) the data is by assigning them different prior probabilities.[1] But, as I argued in chapter 3, prior probabilities, except in highly special cases, are matters of personal, subjective choice—threatening to lead to the relativism we are being challenged to avoid (inviting a MUD-slide, one might say).

Summary of the Strategy to Be Developed

How does appealing to the notion of severity help? While there are many different conceptions of severe tests, such accounts, broadly speaking, hold the following general methodological rule:

> Evidence *e* should be taken as good grounds for *H* to the extent that *H* has passed a *severe test* with *e*.

What I want to argue is that the alternative hypothesis objection loses its sting once the notion of severity is appropriately made out.

It is easy to see that the alternative hypothesis objection instantiated for a method of severe testing *T* is more difficult to sustain than when it is waged against mere entailment or instantiation accounts of inference. The charge of methodological underdetermination for a given testing method, which I equate with the alternative hypothesis objection, must show that *for any evidence test* T *takes as passing hypothesis* H *severely, there are always rival hypotheses that* T *would take as passing equally severely.* While MUD gets off the ground when hypothesis appraisal is considered as a matter of some formal or logical relationship

[1] Indeed, if two hypotheses entail the evidence, then the only way they can be differently confirmed by that evidence by Bayes's theorem is if their prior probability assignments differ.

between evidence or evidence statements and hypotheses, this is not so in our experimental testing framework.

The cornerstone of an experiment is to do something to *make* the data say something beyond what they would say if one passively came across them. The goal of this active intervention is to ensure that, with high probability, erroneous attributions of experimental results are avoided. The error of concern in passing *H* is that one will do so while *H* is not true. Passing a severe test, in the sense I have been advocating, counts for hypothesis *H* because it corresponds to having good reasons for ruling out specific versions and degrees of this mistake.

Stated simply, *a passing result is a severe test of hypothesis* H *just to the extent that it is very improbable for such a passing result to occur, were* H *false.* Were *H* false, then the probability is high that a more discordant result would have occurred. To calculate this probability requires considering the probability a given procedure has for detecting a given type of error. This provides the basis for distinguishing the well-testedness of two hypotheses—despite their both fitting the data equally well. Two hypotheses may accord with data equally well but nevertheless be tested differently by the data. The data may be a better, more severe, test of one than of the other. The reason is that the procedure from which the data arose may have had a good chance of detecting one type of error and not so good a chance of detecting another. What is ostensibly the same piece of evidence is really not the same at all, at least not to the error theorist.

This underscores a key difference between the error statistics approach and the Bayesian approach. Recall that for the Bayesian, if two hypotheses entail evidence *e*, then in order for the two hypotheses to be differently confirmed there must be a difference in their prior probabilities. In the present approach, two hypotheses may entail evidence *e*, while one has passed a far more severe test than the other.

6.2 THE (ERROR)-SEVERITY REQUIREMENT

The general requirement of, or at least preference for, severe tests is common to a variety of approaches (most commonly testing approaches), with severity taking on different meanings. To distinguish my notion from others, I will sometimes refer to it as *error-severity.*

a. *First requirement:* e *must "fit"* H. Even widely different approaches concur that, minimally, for *H* to pass a test, *H* should agree with or in some way fit with what is expected (or predicted) according to *H.* We can apply and contrast our definition with that of other approaches by allowing "*H* passes with *e*" to be construed in many ways (e.g., *H* is supported, *e* is more probable on *H* than on not-*H*, *e* is far from the denial of *H* on some distance measure, etc.[2]). Minimally, *H* does not fit *e* if *e* is improbable under *H.*

b. *Second requirement:* e*'s fitting* H *must constitute a good test of* H. Those who endorse some version of the severity requirement concur that a genuine test calls for something beyond the minimal requirement that *H* fits *e.* A severity requirement stipulates what this "something more" should be.

Following a practice common to testing approaches, I identify "having good evidence (or just having evidence) for *H*" and "having a good test of *H.*" That is, to ask whether *e* counts as good evidence for *H*, in the present account, is to ask whether *H* has passed a good test with *e.* This does not rule out quantifying the goodness of tests.[3] It does rule out saying that "*e* is a poor test for *H*" and, at the same time, that "*e* is evidence for *H.*"

c. *The severity criterion (for experimental testing contexts).* To formulate the pivotal requirement of severe tests, it is sufficient to consider the test outputs—"*H* passes a test *T* with experimental outcome *e*" or "*H* fails a test *T* with experimental outcome *e.*" I am assuming that the

[2] This allows us to state the first requirement for *H* to pass a test with *e* as

a. *H* fits *e*,

with the understanding that a suitable notion of fit, which may vary, needs to be stipulated for the problem at hand. While some accounts of testing construe the fit as logical entailment (with suitable background or initial conditions), except for universal generalizations this is rarely obtained. One way to cover both universal and statistical cases is with a statistical measure of fit, such as *e* fits *H* to the extent that $P(e \mid H)$ is high. (The entailment requirement results in $P(e \mid H)$ being 1.) Because $P(e \mid H)$ is often small, even if *H* is true, passing a test is commonly defined comparatively. Evidence *e* might be said to fit *H* if *e* is more probable under *H* than under all (or certain specified) alternatives to *H.*

There is nothing to stop the hypothesis that passes from being composite (disjunctive). For example, in a Binomial experiment *H* may assert that the probability of success exceeds .6, i.e., $H: p > .6$. The alternative *H'* asserts that $p \leq .6$. In such cases, *e* fits *H* might be construed as *e* is further from alternative hypothesis *H'* than it is from any (simple) member of *H*, where "further" is assessed by a distance measure as introduced in chapter 5.

[3] The question of whether *H*'s passing a test with result *e* provides a good test of *H* may alternately be asked as the question of whether *e* provides confirmation or support for *H.* However, when the question is put this way within a testing approach it should not be taken to mean that the search is for a quantitative measure of degree of support—or else it would be an evidential-relationship and not a testing approach. Rather, the search is for a criterion for determining if a passing result provides good evidence for *H*—although the goodness of a test may itself be a matter of degree.

underlying assumptions or background conditions for a test—whatever they are—are located in the various data models of an experimental inquiry, as delineated in chapter 5. This frees me to characterize the severity requirement by itself. The severity requirement is this:

> *Severity requirement*: Passing a test T (with e) counts as a good test of or good evidence for H just to the extent that H fits e and T is a *severe test* of H,

and the severity criterion (SC) I suggest is this:

> *Severity criterion* 1*a*: There is a very high probability that test procedure T would *not* yield such a passing result, if H is false.

By "such a passing result" I mean one that accords at least as well with H as e does. Its complement, in other words, would be a result that either fails H or one that still passes H but accords less well with H than e does. It is often useful to express SC in terms of the improbability of the passing result. That is:

> *Severity criterion* 1*b*: There is a very low probability that test procedure T would yield such a passing result, if H is false.

One may prefer to state the SC in terms of the measure of accordance or fit. (1*a*) and (1*b*) become

> *Severity criterion* 2*a*: There is a very high probability that test procedure T would yield a worse fit, if H is false.

> *Severity criterion* 2*b*: There is a very low probability that test procedure T would yield so good a fit, if H is false.

While the *a* versions express severity in terms of the test's high probability to detect the incorrectness of H, the equivalent *b* versions express severity in terms of the low probability of its failing to detect the incorrectness of H.

d. *The Severity Criterion in the Simplest Case* (*SC**)*: A Test as a Binomial Statistic.* Standard statistical tests are typically framed in terms of only two possible results: H passes and H fails, although "accept" and "reject" are generally the expressions used rather than "pass" and "fail." This reduction to two results is accomplished by stipulating a cutoff point such that any particular result e that differs from H beyond this cutoff point is classified as failing H; all others pass H. The test, in short, is modeled as a Binomial (or pass-fail) procedure. The severity criterion for this special case is simpler to state than for the general case:

> (SC*) The severity criterion for a "pass-fail" test: There is a very high probability that the test procedure T fails H, given that H is false.[4]

Modeling tests in this "Binomial" manner may be sufficient for specifying a test with appropriate error probabilities. However, it is often too coarse grained for interpreting a particular result, which is why its use leads to many criticisms of standard error statistics—a point to be explained in chapter 11. The trick is to be able to calculate the severity achieved by some *specific outcome* among those the test would take as passing H. That is the reason for my more cumbersome definition.[5] Nevertheless, the severity criterion for the pass-fail (or Binomial) test (SC*), because of its simplicity, is the one I recommend keeping in mind even in arguing from a specific passing result. One need only be clear on how it may be used to arrive at the general SC, the calculation we really want. Let us illustrate.

That H passes with a specific outcome e may be regarded as H having passed with a given *score*, the score it gets with outcome e, just like a score on an exam. Suppose we want to calculate the severity associated with that particular passing score e. We can divide the possible scores into two: scores higher than the achieved score e, and those as low as or lower than e. We have now (re)modeled our test so that it has only two results, and we can apply the simple severity calculation for a pass-fail test. We have

> SC*: The probability is high that test T would *not* yield so high a score for H as e, given that H is false.

Alternatively, in terms of the complement (*b*) we have

> SC*: It is very improbable that H would have passed with so successful a score as e, given that H is false.

We have arrived at the calculation that the more general severity criterion (SC) demands.

How to understand the probabilities referred to in our severity criterion is a question whose answer may be found in the discussion of frequentist probability in the last chapter. A high severity assignment asserts that were we experimenting on a system where H is false, then in a long series of trials of this experiment, it is extremely rare (infrequent) that H would be accorded such a good score; the overwhelming

[4] Calculating SC* considers the probability that an outcome would reach the cutoff for failing H, even if H is false.

[5] This will be clarified further in distinguishing severity from "power" in chapter 11.

preponderance of outcomes would accord *H* a worse fit or a lower score.

Minimum (0) and Maximum (1) Severity

We can get at the commonsense rationale for desiring high severity and eschewing low severity by considering extreme cases of violating or satisfying severity. Here the probabilities of *H* not passing when false may be shown to be 0 and 1 (or practically so), respectively. I begin with the first extreme case, that of a *minimally severe* or a *zero-severity* test.

> *Passing a minimally severe (zero-severity) test: H* passes a zero-severity test with *e* if and only if test *T* would always yield such a passing result even if *H* is false.

In the present account, such a test is no test at all. It has no power whatsoever at detecting the falsity of *H*. If it is virtually impossible for *H* to receive a score less than *e* on test *T*, even if false, then *H*'s receiving score *e provides no reason* for accepting *H;* it fails utterly to discriminate *H* being true from *H* being false.

That a test would always pass *H* even if *H* is false does not entail that *H*'s passing is always erroneous or that *H* is false. *H* may be true. I may even have a warrant for accepting *H*, on *other* grounds. It is just that passing with a zero-severity test itself does not warrant such an acceptance. (That is, one can be right, but for the wrong reasons.)

These remarks accord well with familiar intuitions about whether passing marks on an exam warrant merit of some sort. Consider a test to determine whether a student can recite all the state capitals in the United States; say the hypothesis *H* is that the subject can correctly recite (aloud) all fifty. Suppose that a student passes the test so long as she can correctly assert the capital of any one state. That a person passes this test is not much of a reason to accept *H*, because it is not a very severe test. Suppose now that a student passes the test so long as she can recite *anything* aloud. Granted, being able to recite all fifty capitals entails being able to speak aloud (*H* entails *e*), but this test is even less severe than the first. It is easier (more probable) for a pass to occur, even if the student is *not* able to recite all the state capitals (*H* is false).[6]

Alternatively, if a student passes a test where passing requires reciting all fifty capitals correctly, certainly that is excellent support for hypothesis *H*, that the student can correctly recite them all. This identifies the other extreme, that of a maximally severe test:

[6] Popper (1979, 354) also uses an exam analogy to make this point.

> *Passing a maximally severe (100 percent severity) test: H* passes a maximally severe test with *e* if and only if test *T* would never yield results that accord with *H* as well as *e* does, if *H* is false.

A test is maximally severe if the results that the test takes as passing *H* cannot occur (in trials of the given experimental test), given that hypothesis *H* is false. It is a maximally reliable error probe for *H*. That passing a maximally severe test warrants accepting *H* may seem too obvious to merit noting. After all, in such a test passing with *e* entails *H!* Nevertheless, as will be seen in chapter 8, not all accounts of testing countenance maximally severe tests as good tests.

Let us move from 100 percent severity to merely high severity and see whether the reasoning still holds. Consider two tests, T_1 and T_2.

T_1 is known to have a very high, say a .99, probability of failing a student (giving her an F grade, say) if the student knows less than 90 percent of the material. That is, 99 percent of the time, students ignorant of 10 percent of the material fail test T_1.

Test T_2, let us suppose, is known to have only a 40 percent probability of failing a student who knows less than 90 percent of the material.

T_1 is obviously a more severe test than T_2 in our ordinary use of that term, and likewise in the definition I have given. Passing the more severe test T_1 is good evidence that the student knows more than 90 percent of the material. (For, if she were to know less than 90 percent, test T_1 would, with high probability, .99, have detected this and failed her.) Clearly, all else being equal, better evidence of the extent of a student's knowledge is provided by the report "She passes test T_1," than by the report "She passes test T_2." Passing test T_2 is an altogether common occurrence (probability .6) even if the student knows less than 90 percent of the material.

An Analogy with Diagnostic Tools

Tools for medical diagnoses (e.g., ultrasound probes) offer other useful analogies to extract these intuitions about severity: If a diagnostic tool has little or no chance of detecting a disease, even if it is present (low severity), then a passing result—a clean bill of health—with that instrument fails to provide grounds for thinking the disease is absent. That is because the tool has a very high probability of issuing in a clean bill of health even when the disease is present. It is a highly unreliable error probe. Alternatively, suppose a diagnostic tool has an extremely high chance of detecting the disease, just if present—suppose it to be a highly severe error probe. A clean bill of health with that kind of tool provides strong grounds for thinking the disease is not present. For if the disease were present, our probe would almost certainly have detected it.

It is important to stress that my notion of severity always attaches to a particular hypothesis passed or a particular inference reached. To ask, How severe is this test? is not a fully specified question until it is made to ask, How severe would a test procedure be, if it passed such and such a hypothesis on the basis of such and such data? A procedure may be highly severe for arriving at one type of hypothesis and not another. To illustrate, consider again a diagnostic tool with an extremely high chance of detecting a disease. Finding no disease (a clean bill of health) may be seen as passing hypothesis H_1: no disease is present. If H_1 passes with so sensitive a probe, then H_1 passes a severe test. However, the probe may be so sensitive that it has a high probability of declaring the presence of the disease even if no disease exists. Declaring the presence of the disease may be seen as passing hypothesis H_2: the disease is present. If H_2 passes a test with such a highly sensitive probe, then H_2 has *not* passed a severe test. That is because there is a very low probability of *not* passing H_2 (not declaring the presence of the disease) even when H_2 is false (and the disease is absent). The severity of the test that hypothesis H_2 passes is very low.

Some further points of interpretation are in order.

Severity and Arguing from Error

Experimental learning, I have been saying, may be addressed in a formal or informal mode, although those might not be the best terms. In its formal mode, experimental learning is learning about the probabilities (relative frequencies) of specified outcomes in some actual or hypothetical series of experiments—it is learning about an *experimental distribution*. In its informal mode, experimental learning is learning of the presence or absence of errors. Experimental learning, in its formal mode, is learning from tests that satisfy the severity criterion (SC). In its informal mode, it is learning by means of an argument from error, one variant of which was given in section 3.2. Here are two versions:

> 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[7]) it exists nevertheless detects no error.

> It is learned that an error is present when a procedure of inquiry that has a very high probability of not detecting an error if (and only if) none exists nevertheless detects an error.

That a procedure detects an error does not mean it definitely finds the error. It is generally not known whether the procedure gets it right. It means that a result occurs that the procedure takes as passing the hypothesis that an error is present. An analogous reading is intended for detecting no error.

In the canonical arguments from error, the probabilistic severity requirement may capture the argument from error so well that no distinction between so-called formal and informal modes is needed. In general, however, asserting that a hypothesis *H* passed a highly severe test in this formal sense is but a pale reflection of the actual experimental argument that sustains inferring *H*. The purpose of the formal characterization is to provide a shorthand for the actual argument from error, which necessarily takes on different forms. The formal severity criterion may be seen to represent a systematic way of scrutinizing the appropriateness of a given experimental analysis of a primary question. Referring to the Suppean hierarchy of models from the last chapter, it is a critique at the level of the experimental testing model.

On the one hand, the informal and often qualitative argument from error takes central stage in applying our severity criterion to actual experiments. On the other hand, there are many features of the formal characterization of severity that offer crucial guidance in doing so. This latter point is as important as it is subtle, and to explain it is not as simple as I would wish. Let me try.

In an informal argument from error one asks, How reliable or severe is the experimental procedure at hand for detecting an error of interest? To answer this question, it is essential to be clear about the (probabilistic) properties of the experimental procedure. Our informal thinking about such things may be anything but clear, and formal canonical models (from standard random experiments) may come to the rescue. For example, at the heart of a number of methodological controversies are questions about whether certain aspects of experimental design are relevant to appraising hypotheses. Does it matter whether a hypothesis was constructed to fit the data? Does it matter when we decide how much data to collect? These are two examples that will be taken up in later chapters.

The formal severity criterion, by reminding us that the test procedure may be modeled as a random variable, comes to our aid. For we know that we cannot determine the distribution of a random variable without being clear on what it is that is being taken to vary from trial

[7] The "only if" clause is actually already accommodated by the first requirement of passing a severe test, namely, that the hypothesis fit the data. If the fit required is entailment, then the probability of passing given the hypothesis is true is maximal.

to trial. Is it just the sample mean that varies (e.g., the different proportions of heads in *n* trials)? Or is the very hypothesis that a test procedure winds up testing also varying? Formally modeled (canonical) experiments demonstrate how error probabilities and, correspondingly, severity can be altered—sometimes dramatically—by changing what is allowed to vary. (Doing so is tantamount to changing the question and thereby changing the ways in which answers can be in error.) Carrying a few of the formally modeled test procedures in our experimental tool kit provides invaluable methodological service.

The distinction between the formal model and informal arguments from error also frees us to talk about a hypothesis being true without presuming a realist epistemology. Within a canonical experimental test, the truth of *H* means that *H* gives an approximately correct description or model of some aspect of the procedure generating experimental outcomes. Precisely what this statement of experimental knowledge indicates about the system of interest will vary and will have to be decided on a case by case basis. The main thing to note is that our framework allows numerous interpretations to be given to the correctness of *H*, as well as to the success of *H*. Realists and nonrealists of various stripes can find a comfortable home in error testing. Aside from varying positions on realism, a variety of interpretations of "*H* is true" (and, correspondingly, "evidence indicates that *H* is true") are called for because of the very different kinds of claims that may be gleaned from experiments. (The Kuhnian normal scientist of chapter 2, for example, may view "*H* is true" as asserting that *H* is a satisfactory solution to a normal problem or puzzle.)

Despite this room for diversity, there is uniformity in the pattern of arguments from error. We can get at this uniformity, I propose, by stating what is learned from experiment in terms of the presence or absence of some error (which may often be a matter of degree). For example, a primary hypothesis *H* might be

H: the error is absent,

and not-*H*, that the error is present. (Alternatively, *H* can be construed as denying that it would be an error to assert *H*.) When an outcome is in accordance with *H* and (appropriately) far from what is expected given not-*H*, then the test passes *H*. Error now enters in a second way. The error of concern in passing *H* is that one will do so while *H* is not true, that the error will be declared absent though actually present.

When a test is sufficiently severe, that is, when an argument from error can be sustained, the passing result may be said to be a *good indication of* (or good grounds for) *H*. The resulting knowledge is experimental knowledge—knowledge of the results to be expected were certain experiments carried out.

We now have to tackle the "other hypothesis" objection. For the existence of alternative hypotheses that accord equally well with test results may be thought to strangle any claim purporting that a test's severity is high.

6.3 Is the Other Hypothesis Objection an Objection to Severity?

The thrust of the "other hypothesis" objection is this: the fact that data fit hypothesis *H* fails to count (or to count much) in favor of *H* because the data also fit other, possibly infinitely many, rival hypotheses to *H*. The above characterization of severe tests suggests how this objection is avoidable: mere fitting is not enough! If hypotheses that fit the data equally well were equally well supported (or in some way credited) by the data, then this objection would have considerable weight. But the very raison d'être of the severity demand is to show that this is not so.

Still it might be charged that demanding severity is too demanding. This is Earman's (1992) criticism of me. Examining his criticism allows me to address an anticipated misunderstanding of the severity criterion, namely, the supposition that it requires what I called the Bayesian "catchall" factor (section 4.3).

Earman's Criticism of Error-Severity

In order for hypothesis *H* to pass a severe test, the test must have a low probability of erroneously passing *H*. (This alludes to the *b* forms of SC.) Earman's criticism of my severity requirement seems to be that it requires a low probability to the Bayesian catchall factor. The Bayesian catchall factor (in assessing *H* with evidence *e*), recall, is

$$P(e \mid \text{not-}H).$$

However, satisfying SC does not require computing the Bayesian catchall factor.

The catchall, not-*H*, refers to all possible hypotheses other than *H*, including those that may be conceived of in the future. Assessing the probability of *e* on the catchall requires having a prior probability assignment to the catchall. Assigning a low value to the Bayesian factor on the catchall, while all too easy for a personalist—it is sufficient that he or she cannot think of any other plausible explanation for *e*—is too difficult for a tempered subjectivist or frequentist Bayesian, for it requires, recalling Salmon's remark, that we "predict the future course of the history of science" (Salmon 1991, 329).

Earman grants the *desirability* of a low assignment to the Bayesian catchall factor, because, as we said, the lower its value, the more Bayesian confirmation accrues to *H*. The difficulty he sees is in obtaining it. While I agree that this presents an obstacle for the Bayesian approach to support, satisfying the severity criterion SC does not require computing the Bayesian catchall factor. Because of this, alternatives in the catchall that might also fit the evidence are not the obstacle to obtaining high severity that Earman thinks they are.

Consider the example Earman raises in this connection (I substitute *e* for his *E* to be consistent with my notation):

> If we take *H* to be Einstein's general theory of relativity and *e* to be the outcome of the eclipse test, then in 1918 and 1919 physicists were in no position to be confident that the vast and then unexplored space of possible gravitational theories denoted by ¬GTR does not contain alternatives to GTR that yield the same prediction for the bending of light as GTR. (Earman 1992, 117)

In fact, he continues, there is an endless string of such alternative theories. *The presumption is that alternatives to the GTR that also predict light bending would prevent high severity in the case of the eclipse test.*

But alternatives to the GTR did not prevent the eclipse results from being used to test severely the hypotheses for which the eclipse experiments were designed. Those tests, to be taken up in chapter 8, proceeded by asking specific questions: Is there a deflection of light of about the amount expected under Einstein's law of gravitation? Is it due to gravity? Are alternative factors responsible for appreciable amounts of the deflection? Finding the answers to these questions in a reliable manner did not call for ruling out any and all alternatives to the general theory of relativity.

Take the question of the approximate deflection of light. If this is the primary question of a given inquiry, then alternative answers to it are alternative values of the deflection, not alternatives to the general theory of relativity. If alternative theories predict the same results, so far as the deflection effects go, as Earman says they do, then these alternatives are not *rivals* to the particular hypotheses under test. If the endless string of alternative theories would, in every way, give the same answers to the questions posed in the 1919 tests, then they all agree on the aspects of the gravitation law that were tested. They are not members of the space of alternatives relative to the primary question being addressed.

This reply depends on a key feature of my account of testing, namely, that an experimental inquiry is viewed as a series of models, each with different questions, stretching from low-level theories of data and experiment to higher level hypotheses and theories of interest. In relation to the hypotheses about the deflection effect, alternatives to the general theory of relativity are on a higher level. The higher-level alternatives are not even being tested by the test at hand. Most important, higher-level alternatives pose no threat to learning with severity what they needed to learn in the specific 1919 experiments.

For a silly analogy, consider a dialogue about what can be inferred from an exam (we assume cheating is ruled out):

Teacher: Mary scored 100 percent on my geography final—she clearly knows her geography.

Skeptic: How can you be so sure?

Teacher: Well, it *is* possible that she guessed all the correct answers, but I doubt that any more than once in a million years of teaching would a student do as well as Mary by merely guessing.

Skeptic: But there is an endless string of childhood learning theories that would predict so good a score. Perhaps it's the new text you adopted or our attempts to encourage girls to compete or . . .

Teacher: My final exam wasn't testing any of those hypotheses. They might be fun to test some day, but whatever the explanation of her performance, her score on the final shows me she really knows her geography.

The general lesson goes beyond answering Earman. It points up a strategy for dispelling a whole class of equally good fitting alternatives to a hypothesis *H*. The existence of alternatives at a higher level than *H* is no obstacle to finding high severity for *H*. The higher-level questions, just like the question about the correctness of the whole of the GTR, are simply *asking the wrong question.*

Testing versus Learning About

Saying that the eclipse tests were not testing the full-blown theory of general relativity does not mean that nothing was learned about the theory from the tests. What was learned was the extent to which the theory was right about specific hypotheses, for example, about the parameter λ, the deflection of light.

This points up a key distinction between experimental learning in the present approach and in the Bayesian approach, which may explain why Earman thinks that error-severity founders on the alternative hypothesis objection. For a Bayesian, learning about a hypothesis or theory is reflected in an increase in one's posterior probability assignment to that hypothesis or theory. For a result to teach something about the theory, say the GTR, for a Bayesian, that theory must have received some confirmation or support from that result. But that means the theory, the GTR, must figure in the Bayesian computation. That, in turn, requires considering the probability of the result on the negation of the GTR, that is, the Bayesian catchall factor. That is why Earman's criticism raises a problem for Bayesians.[8]

For the error theorist, in contrast, an experiment or set of experiments may culminate in accepting some hypothesis, say about the existence of some deflection of light. This can happen, we said, if the hypothesis passes a sufficiently severe test. That done, we are correct in saying that we have learned about one facet or one hypothesis of some more global theory such as the GTR. Such learning does not require us to have tested the theory as a whole.

Our approach to experimental learning recommends proceeding in the way one ordinarily proceeds with a complex problem: break it up into smaller pieces, some of which, at least, can be tackled. One is led to break things down if one wants to learn. For we learn by ruling out specific errors and making modifications based on errors. By using simple contexts in which the assumptions may be shown to hold sufficiently, it is possible to ask *one question at a time.* Setting out all possible answers to this one question becomes manageable, and that is all that has to be "caught" by our not-*H*.

Apart from testing some underlying theory (which may not even be in place), scientists may explore whether neutral currents exist, whether dense bodies are real or merely artifacts of the electron microscope, whether F_4 and F_5 chromosomes play any part in certain types of Alzheimer's disease, and so on. In setting sail on such explorations, the immediate aim is to see whether at least one tiny little error can be ruled out, without having to worry about all the ways in which one could ever be wrong in theorizing about some domain, which would only make one feel at sea.

Within an experimental testing model, the falsity of a primary hypothesis *H* takes on a specific meaning. If *H* states that a parameter is greater than some value *c*, not-*H* states that it is less than *c*; if *H* states that factor *x* is responsible for at least *p* percent of an effect, not-*H* states that it is responsible for less than *p* percent; if *H* states that an effect is caused by factor *f*, for example, neutral currents, not-*H* may say that it is caused by some other factor possibly operative in the experimental context (e.g., muons not making it to the detector); if *H* states that the effect is systematic—of the sort brought about more often than by chance—then not-*H* states that it is due to chance. How specific the question is depends upon what is required to ensure a good chance of learning something of interest (much like ensuring satisfaction of the Kuhnian demarcation criterion of chapter 2).

[8] These remarks do not encompass all the ways that the error-severity calculation differs from calculating the Bayesian catchall factor. They simply address the point that was at the heart of Earman's criticism.

I am not denying the possibility of severe tests of higher-level theoretical hypotheses. When enough is learned from piecemeal studies, severe tests of higher-level theories are possible. Kuhn was right that "severity of test-criteria is simply one side of the coin whose other face is a puzzle-solving tradition." The accumulated results from piecemeal studies allow us at some point to say that several related hypotheses are correct, or that a theory solves a set of experimental problems correctly.

Earman (1992, p. 177) discusses for a different reason the progress that has been made in a program by Thorne and Will (1971) to classify theories of gravity, those already articulated as well as other possible theories satisfying certain minimal requirements.[9] Such a program shows which available experiments can eliminate whole chunks of theories (e.g., so-called nonmetric theories of gravity) and which sets of theories are still not distinguished by known experiments, and it indicates how progress might be made by devising experiments to further discriminate between them (e.g., making cosmological observations). Something like this kind of program of partitioning and eliminating chunks of theories is what the present program would call for at the level of large-scale theories.

Much more work is also needed to show how learning in large-scale inquiries proceeds by piecemeal canonical questions. Later I will focus on specific cases, but the philosopher of experiment's search is not for a uniform analysis of high-level testing—at least not in the way that has ordinarily been understood. Still, there are some general strategies for getting at larger questions by inquiring, piecemeal, into smaller errors: testing parameters, estimating the effects of background factors, distinguishing real effect from artifact, and so on with the other canonical errors. There are also general methodological rules for specifying an experimental test from which one is likely to learn, based on

[9] The aim of the program Earman describes is "the exploration of the possibility space, the design of classification schemes for the possible theories, the design and execution of experiments, and the theoretical analysis of what kinds of theories are and are not consistent with what experimental results" (Earman 1992, 177).

background knowledge of the types of errors to be on the lookout for, and strategies for attaining severe tests with limited information.

My concern just now is to get small again, to proceed with some standard tools for severe tests in the experimental models laid out in chapter 5. While they may enable us to take only baby steps, they enable us to take those baby steps severely. Such baby steps are at the heart of the experimenter's focus on what we variously referred to as "topical hypotheses" (Hacking) and "normal puzzles" (Kuhn). Moreover, what these baby steps accomplish will be sufficient for the problem of this chapter: methodological underdetermination. For the reason that arguments about evidence underdetermining hypotheses appear to go through is that we have not bothered to be very clear about what specific evidence is being talked about, what specific hypotheses are being tested, and what specific models of experiment and data are available to constrain inferences.

6.4 Calculating Severity

To determine what, if anything, is learned from an experimental result, we must ask, What, if anything, has passed a severe test? Consider our Binomial experiment for the tea-tasting example (example 5.2, section 5.4). We would pass hypothesis H'—that the probability of successfully discriminating the order of tea and milk in an infusion, p, exceeds .5—by failing or rejecting the null hypothesis H_0: $p = .5$ (i.e., the lady is merely guessing). Here, notice that H_0 is the denial of the hypothesis H' that we pass.

The question concerned the population parameter p. The possible answers, the parameter space, consists of all the possible proportions for p from 0 to 1, but the question asked divides it into two spaces, $p =$ .5 and $p >$ .5. A different inquiry might have tested H_0 against a specific alternative, say $p >$ .8. With minor modifications (of test specification[10]), this calls for the same basic test as in our original partition. So that is a good place to start. It is just this kind of rough and simple question that provides a standard for distinguishing between experimental effects and backgrounds.

In this test the tea-tasting lady scored 60 percent successes in 100 trials. That is, the distance (in the positive direction) between the observed proportion or relative frequency of successes and the hypothesized proportion of successes (.5) equals 2, in standard deviation units (one standard deviation being .05). We ask ourselves: Suppose we were to pass H' (assert that she does better than chance) whenever the experiment results in 60 percent or more successes. How severe would that test procedure be? Would it often lead to mistaking chance effects for systematic or "real" ones?

The test procedure can be written in several ways. One is

> *Test procedure T* (in Binomial experiment 5.2): Pass hypothesis H': $p > .5$) (fail H_0) if at least 60 percent out of the (100) trials are successful.

We then ask the above questions more formally in terms of our "significance question": What is the probability of the experiment producing so large a difference from what is expected under the null hypothesis H_0, if in fact the null hypothesis is true? The answer, we said, was .03—quite easily calculated using the Normal distribution.[11] We have

$$P(\text{test } T \text{ passes } H', \text{ given that } H' \text{ is false } [H_0 \text{ is true}]) = .03.$$

This is the probability of erroneously passing H': the *b* variant of the severity criterion. The state of affairs "such a passing result would *not* have occurred" refers to all of the (100-fold) experimental trials that result in less than 60 percent successes. The probability of this event is 1 minus the probability of erroneously passing H', namely, .97. So the severity for H' is high (.97). This means that in a series of repetitions of the experiment (each with 100 trials), 97 percent of the outcomes would yield less than 60 percent successes, were we in fact experimenting on a population where the probability of success was only .5. We can picture this as the area under the Normal curve to the left of .6, assuming the null hypothesis H_0 to be true (fig. 6.1). By rejecting the null hypothesis H_0 only when the significance level is low, we automatically ensure that any such rejection constitutes a case where the nonchance hypothesis H' passes a severe test. Such a test procedure T can be described as follows:

> *Test Procedure T:* Pass H' whenever the statistical significance level of the difference (between $\bar{X}$ and H_0) is less than or equal to α (for some small value of α).[12]

[10] In this case we would need to increase the sample size beyond 100 to take a rejection of the null hypothesis as severely indicating $p >$ .8. I return to such considerations in chapter 11.

[11] A standard chart on the Normal distribution tells us that a sample mean exceeds the population mean by as much as 2 standard deviations less than 3 percent of the time. The central limit theorem ensures that the Normal approximation is more than adequate.

[12] Because of the adequacy of using the Normal approximation it does not matter if we use "less than" or "less than or equal to α." That is because it is a continuous distribution.

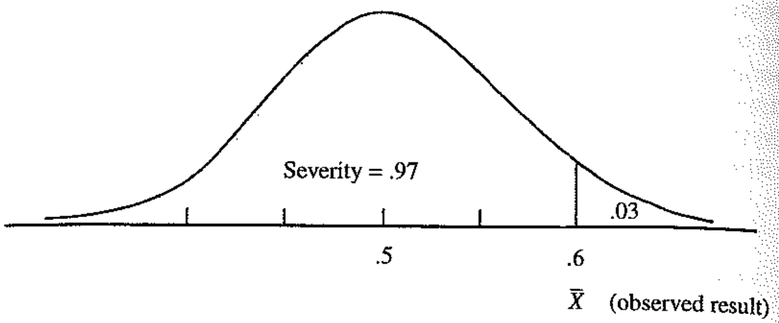


FIGURE 6.1. The severity for passing H' with $\overline{X} = .6$ equals the probability that test T would yield a result closer to H_0 (i.e., .5) than .6 is, given that H_0 is true.

Calculating severity means calculating 1 minus the probability of such a passing result, when in fact the results are due to chance, that is, when H_0 is true. By definition,

> P(T yields a result statistically significant at a level $\leq \alpha$, given that H_0 is true) $= \alpha$.

So the severity of the test procedure T for passing H' is $1 - \alpha$.

As might be expected, were the observed success frequency even higher than 60 percent—say she scored 70 percent or 80 percent successes—the severity for H' would be even higher than .97. Here it is enough to see that the severity of passing H' with result .6 (the 2-standard-deviation cutoff) gives a minimum boundary for how severe the test is for H', and that minimum boundary is high. What is indicated in affirming the nonchance hypothesis H' is that the effect is systematic, that the subject's pattern of correct discernments is not the sort typically brought about just by guessing. Granted, to learn this is typically just a first step in some substantive inquiry. Having found a systematic effect, subsequent questions might be: How large is it (perhaps to subtract it out from another effect)? What causes it? and so on. The aim just now was to illustrate how the current framework allows splitting off one question at a time.

Calculating Severity with Infinitely Many Alternatives

In the Binomial experiment above, the hypothesis that passes the test (the nonchance hypothesis H') had only one alternative hypothesis (the "guessing" hypothesis H_0: $p = .5$).[13] In many cases there are several, even infinitely many, alternatives to the primary hypothesis H for which severity is being calculated. In those cases the "not-H" is a disjunction of hypotheses, H_1 or H_2 or H_3 or. . . . How, it will be asked, can severity be made high in such cases? How can we assess the probability that H does not pass, given either H_1 or H_2 or H_3 or . . . ?

The probability of an outcome conditional on a disjunction of alternatives is not generally a legitimate quantity for a frequentist. Its calculation requires a prior probability assignment to each hypothesis H_i. Lest readers declare, "Aha, you are being Bayesian after all!" I had better explain what SC requires in such cases. It requires that the severity be high against *each such alternative*. In other words, the minimum severity against each of these alternative hypotheses needs to be high (the maximum error probability needs to be low), and prior probability assignments are not required to calculate them.

Consider testing the value of a discrete or continuous parameter μ. Specific examples that will arise later are the mean value for Avogadro's number (chapter 7) and the mean deflection of light (chapter 8). The hypothesis H: μ exceeds some value μ', has as its alternative the complex hypothesis made up of all the values less than μ'. That is, "H is false" here means that μ is one of the infinitely many values less than or equal to μ'. Consider the highest of these values, the one that just makes H false, namely, μ'. This corresponds to the simple alternative hypothesis H': μ equals μ'. The probability that the test would *not* pass H, given that this highest valued alternative, H', were true, is calculated in the usual way. A good test (one with a sensible measure of distance from H) yields even higher severity values for each of the alternative μ values less than μ'. In other words, in a good test, if the test has a high chance of detecting that H is just barely false, it has an even higher chance of detecting that H is even more false. This allows us to say that the severity is high against all alternatives. Good error statistical tests provide just such guarantees. Actual experiments can and often do take their lead from these canonical tests. I return to this in chapter 11.

It may be objected that with substantive questions all the possible alternative hypotheses cannot be set out in the manner of alternative values of a parameter. We may not even know what they are. Even where this is so, it does not present an insurmountable obstacle to experimental testing. In such cases what often may be managed is to find a more general or less precise hypothesis such that when *it* is severely

[13] That H_0, the null hypothesis, plays the role of the alternative here should cause no confusion despite the fact that it is H' that would generally be called the "alternative" in formal statistics. When calculating the severity of a test that passes the non-null hypothesis H', the alternative to H' is the null hypothesis H_0.

tested, there are at the same time grounds for rejecting all the alternatives in a manner meeting the severity requirement. The idea is to partition the possible alternatives to learn about the features that any severely tested hypothesis will affirm. What we try to do, in short, is emulate what is possible in canonical experimental tests.

I can rule out the killer's being over six-feet tall without scrutinizing all six-footers. A single test may allow ruling out all six-footers. Using a similar strategy Jean Perrin was able to rule out, as causes of Brownian motion, all factors outside a certain liquid medium. He did so by arguing that if the observed Brownian motion were due to such external factors—*whatever they might be*—the motion of Brownian particles would follow a specified coordinated pattern. His experimental tests, Perrin argued, would almost surely have detected such a pattern of coordination, were it to exist; but only uncoordinated motion was found. In this way, a whole set of extraneous factors was ruled out. This example will be explored in chapter 7.

6.5 Using Severity to Avoid Misinterpretations of Statistical Tests: The Case of Negative Results

I intend the severity criterion to provide a way of scrutinizing what has been learned from applying standard statistical tests. This scrutiny allows us to go beyond merely following the standard conventions of, say, rejecting the null hypothesis on the basis of a statistically significant result. As such, assessing severity is a tool for avoiding common misinterpretations of standard error statistics. Severity considerations, we can say, serve the "metastatistical" function of scrutinizing error statistical results. They can be used to develop standard tools for avoiding canonical mistakes of interpretation (see section 11.6). These mistakes run to type.

A type of mistake particularly appropriate to consider in the present context concerns statistically *insignificant* results—that is, results where the null hypothesis is *not* rejected by the conventional significance test. A classic flaw we need to be on the lookout for in interpreting such "negative" results is the possibility that the test was not sensitive (severe) enough to detect that H_0 was in error in the first place. (The test, it would be said, has too low a *power.*) In that case, just because the test detects no error with hypothesis H_0 is no indication that the error really is absent. I discuss this well-known error in detail elsewhere (e.g., Mayo 1983, 1985a, 1985b, 1989, 1991b), and will address it in discussing statistical tests in chapter 11. Here my concern is to tie our handling of this canonical error with our avoidance of the alternative hypothesis objection. In particular, it will become clear that a hypothesis (e.g., a null hypothesis of no difference) may accord quite well with data (e.g., a negative result) and yet be poorly tested and poorly warranted by that data.

Learning from Failing to Find a Statistically Significant Difference: The Case of the Pill

A good example is offered by the randomized treatment-control trial on birth-control pills, sketched in example 5.1 (section 5.2). The question of interest concerned parameter Δ, the difference in rates of clotting disorders among a population of women. The question in this study concerned the error of supposing H_0 (no increased risk) when in fact H' is true—there is a positive increase in risk. That is, we tested H_0: $\Delta = 0$ against H': $\Delta > 0$.

The actual difference observed in the Fuertes study was not statistically significant. In fact, the (positive) difference in disease rates that was observed has a statistical significance level of .4. That is to say, 40 percent of the time a difference as large as the one observed would occur even if the null hypothesis is true. (See note 14.)

However, failing to find a statistically significant difference with a given experimental test is not the same as having good grounds for asserting that H_0 is true, that there is a zero risk increase. The reason is that statistically insignificant differences can frequently result even in studying a population with positive risk increases. The argument from error tells us that we may not declare an error absent if a procedure had little chance of finding the error even if it existed. The severity requirement gives the formal analog of that argument.

Of course, the particular risk increase (Δ value) that is considered *substantively* important depends on factors quite outside what the test itself provides. But this is no different, and no more problematic, than the fact that my scale does not tell me what increase in weight I need to worry about. Understanding how to read statistical results, and how to read my scale, informs me of the increases that are or are not indicated by a given result. That is what instruments are supposed to do. Here is where severity considerations supply what a textbook reading of standard tests does not.

Although, by the severity requirement, the statistically insignificant result does not warrant ruling out any and all positive risk increases—which would be needed to affirm H_0: $\Delta = 0$—the severity requirement does direct us to find the smallest positive increase that *can* be ruled out. It directs us to find the value of Δ' that instantiates the following argument:

> *Arguing from a statistically insignificant increase:* Observing a statistically insignificant (positive) difference only indicates that the actual population increase, Δ, is less than Δ' if it is very probable that the test would have resulted in observing a more significant difference, were the actual increase as large as Δ'.

The "if clause" just says that the hypothesis asserting that Δ is less than Δ' must have passed a severe test.

Using what is known about the probability distribution of the statistic here (the difference in means), we can find a Δ value that would satisfy the severity requirement. Abbreviate the value that is found as Δ^*. The above argument says that the statistically insignificant result indicates that the risk increase is not as large as Δ^*. That is, the hypothesis that severely passes, call it H^*, is

H^*: Δ is less than Δ^*.

Let RI be the statistic recording the observed risk increase (the positive difference in disorder rates among treated and untreated women). Then the severity requirement is satisfied by setting $\Delta^* = RI + 2$ standard deviations (estimated). For example, suppose that the particular risk increase is some value RI_{obs} and that this result is not statistically significant. Then the hypothesis

H^*: Δ is less than $RI_{obs} + 2$ standard deviations

would pass a severe test with RI_{obs}.[14] The severity is .97.

Notice that the test result severely *rules out* all increases in excess of Δ^* (i.e., all smaller values pass severely). It thereby illustrates the circumstance discussed in the last subsection—how severely ruling out one hypothesis may entail severely ruling out many others as well. (I return to this example and the question of interpreting statistically insignificant results in chapter 11.)

6.6 SEVERITY IN THE SERVICE OF ALTERNATIVE HYPOTHESIS OBJECTIONS

The standard examples of the last two sections have shown both how to obtain and how to argue from high severity. These standard examples, I believe, let us make short work of the variants of the alternative hypothesis objection. For starters, these cases demonstrate the point I made in grappling with Earman's criticism in section 6.3. We can avoid pronouncing as well tested a whole class of hypotheses that, while implying (or in some other way fitting) a given result, are nevertheless not part of the hypothesis space of the primary test. They are simply asking after the *wrong question,* so far as the given test is concerned.

[14] In the Fuertes et al. (1971) experiment, 9 of the approximately 5,000 treated and 8 of the approximately 5,000 untreated women showed a particular blood-clotting disorder at the end of the study. The observed difference is 1/5,000. For a discussion, see Mayo 1985b.

Alternatives That Ask the Wrong Question

Regarding the Binomial experiment on the tea-tasting lady, examples of wrong question hypotheses would be the variety of hypotheses that might be adduced to explain how the lady achieves her systematic effect, such as psychophysical theories about sensory discrimination, or paranormal abilities. That these other hypotheses predict the pattern observed does not redound to their credit the way the results count in favor of H', that she does better than guessing. This shows up in the fact that they would not satisfy the severity criterion.

The procedure designed to test severely whether the effect is easily explained by chance is not automatically a reliable detector of mistakes about the effect's cause. With regard to questions about the cause of a systematic effect, a whole different set of wrong answers needs to be addressed. At the same time, the existence of these alternative (causal) hypotheses do not vitiate the severity assignment regarding hypotheses for which the test is well designed.

This same argument can be made quite generally to deal with alternatives often adduced in raising the alternative hypothesis objection. While these alternatives to a hypothesis H also fit or accord with the evidence, they may be shown to be less well tested than is H. Often these alternatives are at a higher level in the hierarchy than the primary hypothesis under test (e.g., hypotheses about parameter values when the primary question is about a correlation, questions about the direction of a cause when the primary question is about the existence of a real correlation). There are two main points: First, these alternative hypotheses do not threaten a high severity assignment to the primary hypothesis. Second, it can be shown that these alternatives are not equally severely tested. *Because they ask a different question, the ways in which they can err differ, and this corresponds to a difference in severity.* Moreover, if the primary hypothesis is severely tested, then these alternatives are less well tested. It is not that the nonprimary hypotheses themselves cannot be subjected to other severe tests, although there is certainly no guarantee that they can be. It is simply that they are not tested by the primary test at hand. It follows that hypotheses that entail well-tested hypotheses need not themselves be well tested.

A scientific inquiry may involve asking a series of different primary questions, and each will (typically) require its own hierarchy of experimental and data models. One cannot properly scrutinize hypotheses in isolation from the specific framework in which they are tested.

Alternative Primary Hypotheses

It will be objected that I have hardly answered the alternative hypothesis objection when it becomes most serious: the existence of alternative hypotheses to the primary hypothesis. This is so. But we can handle such cases in much the same fashion as the previous ones—via a distinction in severity.

One point that bears repeating is that I am not aiming to show that all alternatives can always be ruled out. Experimental learning is never guaranteed. What I do claim to show, and all that avoiding MUD requires, is that there are not always equally well tested alternatives that count as genuine rivals, and that there are ways to discriminate hypotheses on grounds of well-testedness that get around alternative hypothesis objections.

Maximally Likely Alternatives. A type of alternative often adduced in raising the alternative hypothesis objection is one constructed after the fact to perfectly fit the data in hand. By perfectly fitting the data, by entailing them, the specially constructed hypothesis H makes the data maximally probable (i.e., $P(e \mid H) = 1$). Equivalently, e makes H *maximally likely.* The corresponding underdetermination argument is that for any hypothesis H there is a maximally likely alternative that is as well or better tested than H is.

The "curve-fitting problem" is really an example of this: for any curve connecting sample points, infinitely many other curves connect them as well. (The infamous Grue problem may be seen as one variant.) The problem of maximally likely alternatives was also a central criticism of the account of testing that Ian Hacking championed in Hacking 1965.[15] In this account, evidence e supports hypothesis H_1 more than hypothesis H_2 if e is more probable given H_1 than given H_2.[16] The trouble is, as Barnard (1972) pointed out, "there *always* is such a rival hypothesis, *viz.* that things just had to turn out the way they actually did" (p. 129).

A classically erroneous procedure for constructing maximally likely rivals: gellerization. Clearly, we are not impressed with many maximally likely hypotheses adduced to explain given evidence, but the challenge for an account of inference is to provide a general and satisfactory way of marking those intuitively implausible cases. Bayesians naturally appeal to prior probabilities, and for reasons already addressed this is unsatisfactory to us. Moreover, at least from the present point of view, this misdiagnoses the problem. The problem is not with the hypothesis itself, but with the unreliability (lack of severity) of the test procedure as a whole. Infamous examples—both formal and informal—serve as canonical cases of how maximally or highly likely hypotheses can be arrived at in ways that yield tests with low or minimal severity. I call them "gellerized" hypothesis tests.

An informal example is that of the Texas sharpshooter. Having shot several holes into a board, the shooter then draws a target on the board so that he has scored several bull's-eyes. The hypothesis, H, that he is a good shot, fits the data, but this procedure would very probably yield so good a fit, even if H is false. A formal variant can be made out with reference to coin-tossing trials:

Example 6.1: A gellerized hypothesis test with coin-tossing trials. The experimental result, let us suppose, consists of the outcomes of n coin-tossing trials–where each trial yields heads or tails. Call the outcome heads a success and tails a failure. For any sequence of the n dichotomous outcomes it is possible to construct a hypothesis after the fact that perfectly fits the data. The primary hypothesis here concerns the value of the parameter p—the probability of success on each coin-tossing trial. The standard null hypothesis H_0 is that the coin is "fair"—that p is equal to .5 on each coin-tossing trial. Thus any alternative hypothesis about this parameter can be considered an alternative primary hypothesis. In any event, this is what our imaginary alternative-hypothesis challenger alleges.

Let $G(e)$ be some such hypothesis that is constructed so as to perfectly fit data e. ($G(e)$ is constructed so that $P[e \mid G(e)] = 1$.) Suppose that $G(e)$ asserts that p, the probability of success, is 1 just on those trials that result in heads, and 0 on the trials that result in tails.[17] It

[15] It was one of the reasons he came to reject the account. See, for example, Hacking 1972.

[16] The probability of e given H_1, $P(e \mid H_1)$, is called the likelihood of H_1. So Hacking's rule of support can also be stated as e supports H_1 more than H_2 if the likelihood of H_1 exceeds the likelihood of H_2.

[17] For example, suppose that e, the result of four tosses of a coin, is heads, tails, tails, heads. That is, $e = s,f,f,s$ where s,f are the abbreviations for "success" and "failure," respectively. Then $G(e)$ would be: the probability of success equals 1 on trials one and four, 0 on trials two and three. The null hypothesis, in contrast, asserts that the probability of success is .5 on each trial. Another G hypothesis that would do the job would assert that the observed pattern of successes and failures will always recur in repeating the n-fold experiment. I owe this second example to I. J. Good.

matters not what if any story accompanies this alternative hypothesis. This hypothesis $G(e)$ says that

> $G(e)$: p equals 1 on just those trials that were successes, 0 on the others.

The test procedure, let us suppose, is to observe the series of however many trials, find a hypothesis $G(e)$ that makes the result e maximally probable, and then pass that hypothesis. In passing $G(e)$, the test rejects the null hypothesis H_0 that the coin is fair.

> *Test procedure T* (in example 6.1): Fail (null) hypothesis H_0 and pass the maximally likely hypothesis $G(e)$ on the basis of data e.

The particular hypothesis $G(e)$ erected to perfectly fit the data will vary in different trials of our coin-tossing experiment, but for every data set, some such alternative may be found. Therefore, any and all experimental results are taken to fail null hypothesis H_0 and pass the hypothesis $G(e)$ that is constructed to fit data e—even when $G(e)$ is false and H_0 is true (i.e., even when the coin is "fair"). In a long-run series of trials on a fair coin, this test would always fail to correctly declare the coin fair. Hence the probability of passing $G(e)$ erroneously is maximal—the severity of this test procedure is minimal.

To calculate severity in cases where the hypothesis is constructed on the basis of data e, it is important to see that two things may vary: the hypothesis tested as well as the value of e. One must include, as part of the testing procedure, the particular rule that is used to determine which hypothesis to test. When the special nature of this type of testing procedure is taken into account, our severity criterion SC becomes

> *SC with hypothesis construction:* There is a very high probability that test procedure T would *not* pass the hypothesis it tests, given that the hypothesis is false.

To ascertain whether SC is satisfied, one must consider the particular rule for designating the hypothesis to test.

Let the test procedure T be the one just described. The hypothesis that T tests on the basis of outcome e is $G(e)$. There is no probability that test T would *not* pass $G(e)$, even if $G(e)$ were false. Hence the severity is minimal (i.e., 0). In other words, the test procedure T is a maximally unreliable probe when it comes to detecting the relevant error (the error of rejecting H_0 when H_0 is true). This amounts to the defining characteristic of what I call a *gellerized hypothesis*—or, more precisely, a gellerized hypothesis-testing procedure. With a gellerized procedure,

the hypothesis selected for testing is one that is constructed to provide an excellent fit for the data, but in such a way that the constructed hypothesis passes a test with minimal (or near minimal) severity.

The manner in which the severity criterion eliminates such gellerized alternatives is important for it hinges on the distinctive feature of error statistical approaches—the centrality of error probabilities. How this contrasts with other approaches will become much clearer later (in chapters 8–11). It should be stressed that gellerized hypotheses are deemed poorly tested by my account not because they are constructed after the fact to fit the data. As I shall argue (in chapter 8), such after-trial constructions ("use-constructed" hypotheses) can pass with maximal severity. They are deemed poorly tested because in gellerized constructions the tactic yields a low or 0 severity test. A primary hypothesis that fits the outcome *less* well than this type of rigged alternative may actually be better, more severely, tested.

The example of gellerization (which comes in several forms), then, is a canonical example of a minimally severe test. As with all canonical examples, it is a basis for criticizing substantive cases that while less obviously fallacious are quite analogous.

Practically Indistinguishable Alternatives. What about alternatives that cannot be distinguished from a primary hypothesis H on the grounds of severity because they differ too minutely from H? This occurs, for example, when H is an assertion about a continuous parameter. My quick answer is this: if there are alternatives to H that are substantive rivals—one differing merely by a thousandth of a decimal is unlikely to create a substantive rival—and yet they cannot be distinguished on the grounds of severity, then that is grounds for criticizing the experimental test specifications (the test was insufficiently sensitive). It is not grounds for methodological underdetermination.

Empirically Equivalent Alternatives. We have yet to take up what some might consider the most serious threat to a methodology of testing: the existence of rival primary hypotheses that are empirically equivalent to H, not just on existing experiments but on all possible experiments. In the case where the alternative H' was said to ask the wrong question, it was possible to argue that the severity of a test of primary hypothesis H is untouched. (If the ways in which H can err have been well probed and ruled out, then H passes a severe test. There is no reason to suppose that such a test is any good at probing the ways in which H' can err.) But the kind of case we are to imagine now is not like that. Here it is supposed that although two hypotheses, H and H', give different answers to the same primary question, both have the same testable consequences. Does it follow that a severity assessment is unable to discriminate between any tests they both pass?

That depends. If it is stipulated that any good test is as likely to pass H, although H' is true, as it is to pass H' although H is true—if it is stipulated that any test must have the same error probabilities for both hypotheses—then it must be granted. In that case no severe test can indicate *H as opposed to H'*. The best example is mathematical, the two hypotheses being Euclidean and non-Euclidean geometry. But apart from certain, not entirely uncontroversial cases in physics, there is no reason to suppose such pairs of rivals often exist in science.[18] Moreover, even if we grant the existence of these anomalous cases, this would fail to sustain MUD, which alleges that the problem exists for *any* hypothesis. There is no reason to suppose that every hypothesis has such a rival.

We can go further. When one looks at attempts to argue *in general* for the existence of such empirically (or testably) equivalent rivals, one finds that severity considerations discriminate among them after all. In fact, one finds that such attempts appeal to tactics remarkably similar to those eschewed in the case of gellerized hypotheses. They too turn out to be "rigged" and, if countenanced, lead to highly unreliable test procedures.

Richard Miller (1987) gives a good example in objecting to alleged empirically equivalent, "just-as-good" alternatives. He asks, "What is the theory, contradicting elementary bacteriology, that is just as well confirmed by current data?" (p. 425) Granted, an alternative that can be constructed is "that bacteria occasionally arise spontaneously but only when unobserved." The severe testing theory dismisses such a general tactic the same way it dismisses an alleged parapsychologist's claim that his powers fail when scientists are watching. Such a tactic (gellerization again!) allows the alternative hypothesis to pass the test, but only at the cost of having no (or a low) chance of failing, even if it is false—at the cost of adopting a minimally severe test. I condemn such tests because one cannot learn from them.

But, the alternative hypothesis objector may persist, doesn't the existence of such an alternative prevent a high-severity assignment to the hypothesis of elementary bacteriology? No. The grounds for assessing how severely this hypothesis passes are a separate matter. It

passes a severe test to the extent that there are good grounds for arguing that were the bacteriology hypothesis false, then it almost surely would have been found to be false. Such grounds may or may not exist (in the bacteriology case, as Miller notes, it seems that they do). What matters is that no obstacle to such grounds is presented by a rigged alternative, R. Hypothesis R, in effect, makes the following assertion:

> *Rigged hypothesis R:* a (primary) alternative to H that, by definition, would be found to agree with any experimental evidence taken to pass H.

Consider the general procedure of allowing, for any hypothesis H, that some rigged alternative or other is as warranted as H. Even where H had repeatedly passed highly severe probes into the ways H could err, this general procedure would always sanction the following argument against H: all existing experiments were affected in such a way as to systematically mask the falsity of H. That argument procedure is highly unreliable. It has a very high (if not a maximal) probability of erroneously failing to discern the correctness of H.

Alternatives about Experimental Assumptions

One way of challenging the claim to have severely tested H is by challenging the experimental assumptions. Assigning a high severity to a primary hypothesis H assumes that the experimental assumptions are approximately met. In fact, the key feature of well-specified experimental tests is that the only nonprimary hypotheses that need to be worried about, for the sake of answering the single question at hand, are challenges to the assumptions of the experimental model. Chapter 5 discusses how to handle these assumptions (they were placed at the stage of checking the data models yet lower down in the hierarchy), so a sketch should suffice.

Again, the procedures and style of argument for handling experimental assumptions in the formal, canonical inquiries are good standards for learning in actual, informal experiments. These experimental procedures fall into two main groups. The first consists of the various techniques of experimental design. Their aim is to satisfy experimental assumptions before the trial is carried out. The second consists of procedures for separately testing experimental assumptions after the trial. Often this is done by means of the "same" data used to pass the primary hypothesis, except that the data are modeled differently. (For instance, the same sequence of trials may be used to answer questions about the assumptions of the Binomial experiment—e.g., is the cause of the effect the color of the cups? The data set is remodeled to ask a different

[18] Earman (1993) suggests that the existence even of exotic empirically indistinguishable rivals is enough to make us worry that only a lack of imagination keeps us from recognizing others "all over the map" (p. 31).

question.) With respect to the statistical assumptions of the two tests studied earlier (the pill and tea-tasting experiments), a whole battery of separate statistical tests is available, often with trivial assumptions. Moreover, we know from the central limit theorem (chapter 5) that with such a large sample size (100), the Normal approximation to the Binomial experiment in the tea-tasting test is easily justified without further checks.

Recall that I initially separated out the error of violating experimental assumptions (the fourth canonical error [error *d*] in section 1.5) because in general a far less demanding type of argument is needed here. A rough idea of the distribution of the experimental test statistic suffices to say, approximately, how often it is likely to be further from hypothesized values. A host of virtually assumption-free checks often does the job.

In other cases it may be necessary to generate additional data to rule out possible auxiliary factors, such as when the ceteris paribus conditions become suspect. In yet other cases alternative hypotheses may be rejected on the basis of evidence from earlier experiments, now part of the background knowledge, or because they force inconsistent assignments to physical constants. Chapter 7, on Brownian motion, and the discussion of the eclipse experiments in chapter 8 contain illustrations of both types of strategies.

All of this of course requires astute experimental design and/or analysis. By means of the experimental planning, the logically possible explanations of results are effectively rendered actually or practically impossible. The experimentalist whose aim is to get it right does not appeal to hard cores, prior probabilities, or the like; he or she appeals to the various techniques in the experimentalist's tool kit.

To reiterate, I do not hold that relevant alternatives can always successfully be put to rest in these ways. If the threats cannot be ruled out satisfactorily, then the original argument alleging *H* to be indicated is vitiated. Even so, it does not follow that this alternative hypothesis is itself well tested. To say that the alternative is well tested requires a separate argument showing that *it* has passed a severe test.

6.7 SEVERITY, POPPERIAN STYLE

In appealing to severity to answer the other hypothesis objection, it is clear that the probability in SC does not just fall out from some logical relationships between statements of evidence and hypotheses. We must look at the particular experimental context in which the evidence was garnered and argue that its fitting a hypothesis is very improbable,

if that hypothesis is false. This probability refers to the variable behavior of the test rule in (actual or hypothetical) repetitions of the experiment, and the falsity of the hypothesis refers to the presence of some specific error. This relativity to an experimental testing model and the focus on (frequentist) probabilities of test procedures distinguish my account, particularly from others that likewise appeal to probabilities to articulate the criterion for a good or severe test—even from accounts that at first blush look similar, most notably Popper's.

It is important to distinguish Popperian severity from ours because, like the case of the straight rule, Popperian testing has been successfully criticized as open to the alternative hypothesis objection. I explained in chapter 1 why earning a "best tested so far" badge from Popper would not suffice to earn a "well-tested" badge from me. There are, however, several places in which Popper appears to be recommending the same kind of severity requirement as I am. I suspect that Popper's falsification philosophy is congenial to so many scientists because they suppose he is capturing the standard error-testing principles that are at the heart of experimental practice. Less advertised, and far less congenial, is Popper's negativism, that, as he admits, corroboration yields nothing positive, and that it never warrants relying on well-tested hypotheses for future applications. But Popper's most winning slogans are easily construed as catching the error-severity spirit. Here are a few:

> Mere supporting instances are as a rule too cheap to be worth having; they can always be had for the asking; thus they cannot carry any weight; and 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, 130; emphasis added)

> The theoretician will therefore try his best to detect any false theory . . . he will try to "catch" it. That is, he will . . . try to think of cases or situations in which it is likely to fail, if it is false. Thus he will try to construct *severe* tests, and *crucial* test situations. (Popper 1979, 14)

It is not difficult to hear these passages as echoing the goal of severe tests in the sense of SC. Nevertheless, this goal is not accomplished by means of the logical relationships between evidence and hypothesis that Popper calls for. (The particular mathematical formulas Popper offered for measuring the degree of severity are even more problematic and they will not be specifically considered here.) Popper kept to the logical notion of probability, although no satisfactory account of that concept was ever developed. He failed to take what may be called the "error probability turn."

In the next passage, and elsewhere, Popper describes the type of context that he takes as providing grounds for calling a test severe:

> A theory is tested . . . by applying it to very special cases—cases for which it yields results *different from those we should have expected without that theory, or in the light of other theories* . . . those crucial cases in which we should expect the theory to fail if it is not true. (Popper 1962, 112; emphasis added)

Here Popper plainly states that the reason he thinks that hypothesis *H* can be expected to fail if false is that background and alternative hypotheses predict not-*e*—*e* being the result taken to corroborate *H*. That is to say, for Popper, a nonfalsified hypothesis *H* passes a severe test with *e* if all alternatives to *H* that have so far been considered or tested entail not-*e* (or render not-*e* highly probable). A weaker construal requires only that the alternatives say nothing about whether *e* or not-*e* will occur.

Later we will see that the general question of what counts as a severe test is alternately put in terms of the question of what counts as novel evidence for a hypothesis. The answer given by Popper's requirement here is tantamount to requiring that *e* be novel in the sense Alan Musgrave calls *theoretically novel*. The evidence taken to pass *H* is theoretically novel if it is not already derivable from background theories. Lakatos and Musgrave (at times) endorsed both weak and strong construals:

> According to this [theoretical] view, a new theory is independently testable (or predicts a "novel fact") if it predicts something which is not also predicted by its background theory. Hence there are two kinds of independent or novel predictions, tests of which are severe. . . . First, there are predictions which *conflict* with the predictions of the background theory—tests of these will be crucial tests between the new theory and the old. Second, there are predictions concerning phenomena about which the background theory predicts nothing at all—tests of these will also be independent tests and severe ones. (Musgrave 1974, 15–16)

In this view, the requirement for a predicted fact *e* to count as a severe test of *H* may be understood either in the strong form—*e* disagrees with what would be expected given each alternative *H'*—or in the weak form—*H'* says nothing about *e*. I will continue to take the former, stronger version, as the one Popper champions.[19] That is,

Popperian severity:

1. *H* entails *e* ($P(e \mid H) = 1$) or *e* is very probable given *H*
2. Each available *H'* alternative to *H* counterpredicts *e*.

Since it is not clear whether condition 2 requires that the considered alternatives entail not-*e* or simply that each renders *e* very improbable, saying that *H'* counterpredicts *e* denotes either, except where specifically noted. We might state Popperian severity as follows: *e* is a severe test of *H* if *H* predicts *e*, while *e* is anomalous for all other known alternative hypotheses.

There is no demand that a specific testing context be delineated, there are just these two requirements in terms of the logical relationships between statements of evidence and hypotheses. In contrast with the present account, the relevant hypotheses need not be answers to some primary question; they can be anything at all. It is easy to see how the alternative hypothesis objection gets off the ground. Adolf Grünbaum (1978) gets it off quite well.

Grünbaum's "Alternative Hypothesis" Objection to Popperian Severity

Referring to Popper's statement above (Popper 1962, 112), Grünbaum rightly asks how

> qua *pure* deductivist, can Popper possibly maintain without serious inconsistency, as he does, that *successful* results [of tests severe in his sense] should "count" in favor of the theory . . . in the sense that in these "crucial cases . . . we should expect the theory to fail if it is not true"? (Grünbaum 1978, 130)

Passing a severe test in Popper's sense, Grünbaum claims, would leave "the truth of the 'infinitely' weaker disjunction . . . of ALL and only those hypotheses which individually entail [*e*]" (p. 130).[20] And Popper himself acknowledges the existence of infinitely many alternatives.

In other words, suppose that outcome *e* is observed. The hypotheses that entail not-*e* are rejected, and *H*, which entails *e*, passes the test. True, *H* has not yet been refuted—but neither have the infinitely many not-yet-considered other hypotheses that also entail *e*. Evidence *e*, it

[19] Popper confirmed that this was his view in a private communication. If *H* may be constructed after the data, then so long as *H'* does not entail not-*e* it is easy to see how this condition can be satisfied while the test of *H* nevertheless has low severity. A version of the so-called tacking paradox will serve this function. Simply let *H* = *H'* and *e*. Here *H* perfectly fits *e* (i.e., it entails *e*). But since this can be done for any hypothesis *H'*, such an agreement may be assured whether or not *H* is false. This criticism appears in Worrall 1978, 330.

[20] He asks "of what avail is it to Popper, *qua deductivist*, that by predicting *C*, *H* is one of an infinitude of theories . . . incompatible with those *particular* theories with which scientists had been working by way of historical accident?" (Grünbaum 1978, 131).

seems, counts as much for these other alternatives as it does for *H*. As Grünbaum puts it (using *C* for outcome *e*),

> according to Popper's definition . . . the experiment *E* which yielded the riskily predicted *C* does qualify as a "severe test" of *H*. But surely the fact that *H* makes a prediction *C* which is incompatible with the prior theories constituting the so-called "background knowledge" *B* does *not* justify the following contention of Popper's: A *deductivist* is entitled to expect the experiment *E* to yield a result *contrary* to *C*, *unless H is true*. (P. 131)

For even if *H* is false, its falsity is not weeded out. That is because some true (but not yet considered) hypothesis predicts the same outcome that allows *H* to pass. Given the Popperian definition of severity, and given the assumption that there are always infinitely many hypotheses that entail evidence, Grünbaum's worry is well founded. Nor is the situation ameliorated by the additional requirement Popper often advanced, that the hypotheses precede the data (that the data be "temporally novel," to use a term taken up in chapter 8).[21]

The error-severity requirement, in contrast, exists only as part of an experimental account whose central mission is to create situations and specify procedures where we are entitled to expect the experiment to fail *H*, if *H* is false. It is easy to see that satisfying Popper's severity criterion is not sufficient to satisfy ours. One example will suffice.

Cancer Therapies

Each chemotherapeutic agent hypothesized as being the single-bullet cure for cancer has repeatedly failed to live up to its expectations. An alternate, unorthodox treatment, let us imagine, accords with all the available evidence. Let us even imagine that this nonchemotherapeutic hypothesis predicts the chemotherapeutic agents will fail. As such, it may be accorded (one of) Popper's "well-tested" badges. On the error-severity account, the existing data from tests of chemotherapeutic agents provide no test at all of the alternative treatment, because these tests have not probed the ways in which claims made for this alternative treatment can be in error. To count an alternate cancer treatment as well tested simply because it accords with the results of chemotherapeutic trials is to follow a demonstrably unreliable procedure.

Here are the bare bones of this kind of example: A given hypothesis *H* predicts that there will be a significant correlation between *A* (e.g., a cancer drug) and *B* (the remission rate). Alternative hypothesis *H'* predicts that there will be no significant correlation between factors *A* and *B* in the experimental trials. For example, *H'* may assert that only a new type of laetrile treatment can help. The data, let us suppose, are just what *H'* predicts—no significant correlation between *A* and *B* (e.g., in remission rate). *H'* passes a Popperian severe test, but for the error-tester it has passed no genuine test at all (or at best a very weak one).

My criticism of Popperian severity is not merely that credit cannot accrue to currently passing hypotheses because there are, invariably, not-yet-considered alternatives that also would pass. It is rather that, as we have just seen, a good test is not constituted by the mere fact that a hypothesis fits the evidence counterpredicted by existing alternatives.[22] Often we can go further and argue that the test is poor because it did not guard against the types of errors that needed to be guarded against.

If *H* and *H'* are the only possible alternatives—and *H'* entails not-*e* while *e* occurs—then *e* is a maximally severe test of *H* in our sense (and presumably everyone else's). But, in general, Popperian severity is not sufficient for severity in our sense. Neither is Popperian severity *necessary* for error-severity. To consider Popperian severity necessary for a good test would seem to prevent any data already entailed by a known hypothesis to count as severely testing a new hypothesis. (This is argued in Worrall 1978a, 330–31.)

This should finish up the problem with Popperian corroboration first posed in chapter 1. Corroboration, passing a test severe in Popper's sense, says something positive only in the sense that a hypothesis has not been found false—this much Popper concedes. But Popper also suggests that the surviving hypothesis *H* is the best-tested theory so far. I have argued that Popperian tests do not accomplish this. After all, as Popper himself insists, for a hypothesis to be well tested it must have

[21] For Popper, evidence *e* cannot count as a severe test of *H* if *e* is already explained by other hypotheses. In reality, however, the very newness of a phenomenon may count against the first hypothesis to explain it, because one suspects there may be lots of other untried explanations. Until there is some work on the matter, there are not yet grounds to think that *H* would have failed if it were false. I return to this in chapter 8.

[22] Determining if SC is met by Popper's criterion requires asking, "What is the probability of the conditions for *H*'s passing a Popperian severe test being satisfied (in the case at hand) even if *H* is false?" SC requires two things of any test rule: first, that we be able to approximately determine the probability that it results in *H*'s passing even if *H* is false; and second, that this probability is determined to be low. But Popper's severity condition does not provide grounds for assigning a low probability to erroneously passing *H*.

been put through the wringer. One needs to be able to say that *H* had little chance of withstanding the inquiry if false. It is a mistake to consider a result counterpredicted by known alternatives to *H* as automatically putting *H* through the wringer.

Looking at the problem in terms of the logical relationships between evidence and hypotheses ignores all of the deliberate and active intervention that provides the basis for arguing that if a specific error is committed, it is almost certain to show up in one of the results of a given probe (series of tests). By such active intervention one can substantiate the claim that we should expect (with high probability) hypothesis *H* to fail a given test, if *H* is false. And we can do so even if we allow for the possibility of infinitely many alternative hypotheses.

Granted, arguing that a hypothesis is severely tested takes work. In many cases the most one can do is approximate the canonical cases that ground formal statistical arguments. But often it can be argued that a hypothesis is severely tested—even if it means modifying (weakening) the hypothesis. By deliberate and often devious methods, experimenters are able to argue that the test, in context, is severe enough to support a single answer to a single question.

6.8 My Reply to the Alternative Hypothesis Objection

Let us recapitulate how my account of severe testing deals with alternative hypothesis objections that are thought to be the basis for MUD. The MUD charge (for a method of severe testing *T*) alleges that for any evidence test *T* takes as passing hypothesis *H* severely, there is always a substantive rival hypothesis *H'* that test *T* would regard as having passed equally severely. We have shown this claim to be false, for each type of candidate rival that might otherwise threaten our ability to say that the evidence genuinely counts in favor of *H*. Although *H'* may accord with or fit the evidence as well as *H* does, the fact that each hypothesis can err in different ways and to different degrees shows up in a difference in the severity of the test that each can be said to have passed. The same evidence effectively rules out *H*'s errors—that is, rules out the circumstances under which it would be an error to affirm *H*—to a different extent than it rules out the errors to affirming *H'*.[23]

This solution rests on the chief strategy associated with my experimental testing approach. It instructs one to carry out a complex inquiry by breaking it down into pieces, at least some of which will suggest a question that can be answered by one of the canonical models of error. (In some cases one actually carries out the statistical modeling, in others it suffices to do so informally, in ways to be explained.) With regard to the local hypotheses involved in asking questions about experimental mistakes, the task of setting out all possible answers is not daunting. Although it may be impossible to rule out everything at once, we can and do rule out one thing at a time.

Naturally, even if all threats are ruled out and *H* is accepted with a severe test, *H* may be false. The high severity requirement, however, ensures that this erroneous acceptance is very improbable, and that in future experiments the error will likely be revealed.

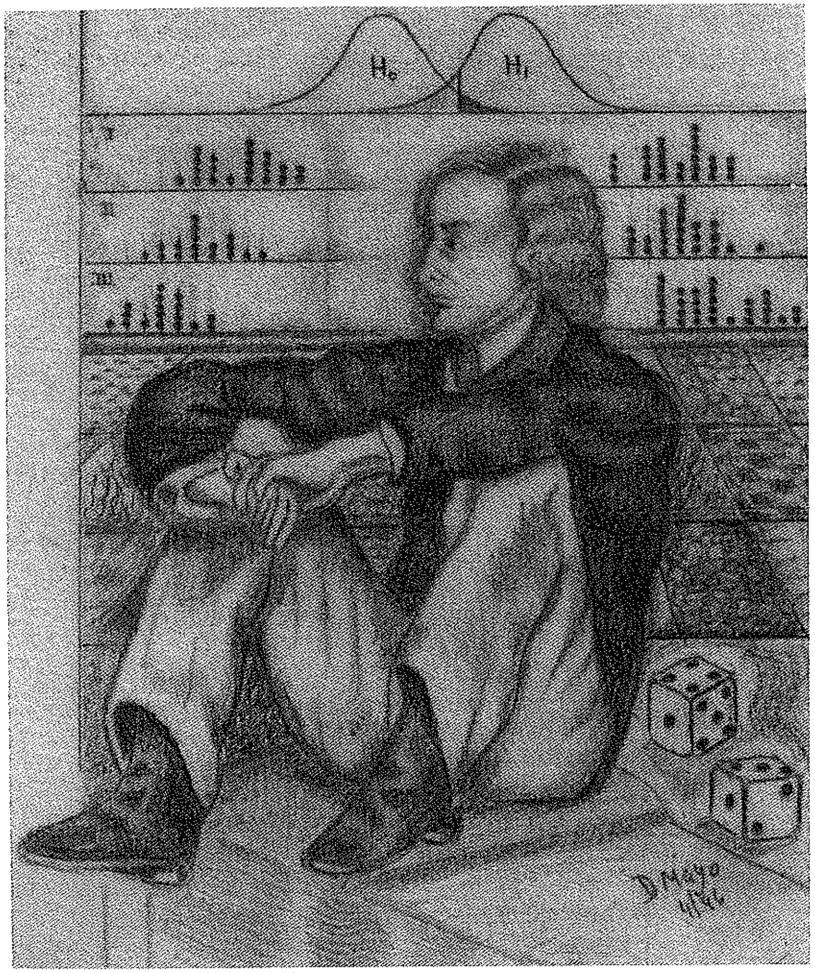
The thrust of experimental design is to deliberately create contexts that enable questions to be asked one at a time in this fashion. In focusing too exclusively on the appraisal of global theories, philosophers have overlooked how positive grounds are provided for local hypotheses, namely, whenever evidence counts as having severely tested them. By attempting to talk about data and hypotheses in some general way, apart from the specific context in which the data and hypothesis are generated, modeled, and analyzed to answer specific questions, philosophers have missed the power of such a piecemeal strategy, and underdetermination arguments have flourished.

Having set out most of the needed machinery—the hierarchy of models, the basic statistical test, and the formal and informal arguments from severe tests—it is time to explore the themes here advanced by delving into an actual scientific inquiry. This is the aim of the next chapter.

[23] This strategy for distinguishing the well-testedness of hypotheses can also be used to resolve the philosophical conundrums known as the Grue paradox and the Ravens paradox.

Science and Its Conceptual Foundations

A series edited by David L. Hull



"I might recall how certain early ideas came into my head
as I sat on a gate overlooking an experimental blackcurrant plot . . . ."
—E. S. Pearson, "Statistical Concepts in Their Relation to Reality"

Deborah G. Mayo

Error and the Growth of Experimental Knowledge

THE UNIVERSITY OF CHICAGO PRESS
Chicago and London